

## SMPL: Simulated Industrial Manufacturing and Process Control Learning Environments

- Mohan Zhang · Xiaozhou Wang · Benjamin Decardi-Nelson · Bo Song · An Zhang · Jinfeng Liu · Sile Tao · Jiayi Cheng · Xiaohong Liu · Dengdeng Yu · Matthew Poon · Animesh Garg
- abstract@[open-review](#): Traditional biological and pharmaceutical manufacturing plants are controlled by human workers or pre-defined thresholds. Modernized factories have advanced process control algorithms such as model predictive control (MPC). However, there is little exploration of applying deep reinforcement learning to control manufacturing plants. One of the reasons is the lack of high fidelity simulations and standard APIs for benchmarking. To bridge this gap, we develop an easy-to-use library that includes five high-fidelity simulation environments: BeerFMTEnv, ReactorEnv, AtropineEnv, PenSimEnv and mAbEnv, which cover a wide range of manufacturing processes. We build these environments on published dynamics models. Furthermore, we benchmark online and offline, model-based and model-free reinforcement learning algorithms for comparisons of follow-up research.

## Avalon: A Benchmark for RL Generalization Using Procedurally Generated Worlds

- Joshua Albrecht · Abraham Fetterman · Bryden Fogelman · Ellie Kitanidis · Bartosz Wróblewski · Nicole Seo · Michael Rosenthal · Maksis Knutins · Zack Polizzi · James Simon · Kanjun Qiu
- abstract@[open-review](#): Despite impressive successes, deep reinforcement learning (RL) systems still fall short of human performance on generalization to new tasks and environments that differ from their training. As a benchmark tailored for studying RL generalization, we introduce Avalon, a set of tasks in which embodied agents in highly diverse procedural 3D worlds must survive by navigating terrain, hunting or gathering food, and avoiding hazards. Avalon is unique among existing RL benchmarks in that the reward function, world dynamics, and action space are the same for every task, with tasks differentiated solely by altering the environment; its 20 tasks, ranging in complexity from eat and throw to hunt and navigate, each create worlds in which the agent must perform specific skills in order to survive. This setup enables investigations of generalization within tasks, between tasks, and to compositional tasks that require combining skills learned from previous tasks. Avalon includes a highly efficient simulator, a library of baselines, and a benchmark with scoring metrics evaluated against hundreds of hours of human performance, all of which are open-source and publicly available. We find that standard RL baselines make progress on most tasks but are still far from human performance, suggesting Avalon is challenging enough to advance the quest for generalizable RL.

## CLEVRER-Humans: Describing Physical and Causal Events the Human Way

- Jiayuan Mao · Xuelin Yang · Xikun Zhang · Noah Goodman · Jiajun Wu
- abstract@[open-review](#): Building machines that can reason about physical events and their causal relationships is crucial for flexible interaction with the physical world. However, most existing physical and causal reasoning benchmarks are exclusively based on synthetically generated events and synthetic natural language descriptions of the causal relationships. This design brings up two issues. First, there is a lack of diversity in both event types and natural language descriptions; second, causal relationships based on manually-defined heuristics are different from human judgments. To address both shortcomings, we present the CLEVRER-Humans benchmark, a video reasoning dataset for causal judgment of physical events with human labels. We employ two techniques to improve data collection efficiency: first, a novel iterative event cloze task to elicit a new representation of events in videos, which we term Causal Event Graphs (CEGs); second, a data augmentation technique based on neural language generative models. We convert the collected CEGs into questions and answers to be consistent with prior work. Finally, we study a collection of baseline approaches for CLEVRER-Humans question-answering, highlighting great challenges set forth by our benchmark.

## MOMA-LRG: Language-Refined Graphs for Multi-Object Multi-Actor Activity Parsing

- Zelun Luo · Zane Durante · Linden Li · Wanze Xie · Ruochen Liu · Emily Jin · Zhuoyi Huang · Lun Yu Li · Jiajun Wu · Juan Carlos Niebles · Ehsan Adeli · Fei-Fei Li
- abstract@[open-review](#): Video-language models (VLMs), large models pre-trained on numerous but noisy video-text pairs from the internet, have revolutionized activity recognition through their remarkable generalization and open-vocabulary capabilities. While complex human activities are often hierarchical and compositional, most existing tasks for evaluating VLMs focus only on high-level video understanding, making it difficult to accurately assess and interpret the ability of VLMs to understand complex and fine-grained human activities. Inspired by the recently proposed MOMA framework, we define activity graphs as a single universal representation of human activities that encompasses video understanding at the activity, sub-activity, and atomic action level. We redefine activity parsing as the overarching task of activity graph generation, requiring understanding human activities across all three levels. To facilitate the evaluation of models on activity parsing, we introduce MOMA-LRG (Multi-Object Multi-Actor Language-Refined Graphs), a large dataset of complex human activities with activity graph annotations that can be readily transformed into natural language sentences. Lastly, we present a model-agnostic and lightweight approach to adapting and evaluating VLMs by incorporating structured knowledge from activity graphs into VLMs, addressing the individual limitations of language and graphical models. We demonstrate strong performance on few-shot activity parsing, and our framework is intended to foster future research in the joint modeling of videos, graphs, and language.

## CEDe: A collection of expert-curated datasets with atom-level entity annotations for Optical Chemical Structure Recognition

- Rodrigo Hormazabal · Changyoung Park · Soonyoung Lee · Sehui Han · Yeonsik Jo · Jaewan Lee · Ahra Jo · Seung Hwan Kim · Jaegul Choo · Moontae Lee · Honglak Lee
- abstract@[open-review](#): Optical Chemical Structure Recognition (OCSR) deals with the translation from chemical images to molecular structures, this being the main way chemical compounds are depicted in scientific documents. Traditionally, rule-based methods have followed a framework based on the detection of chemical entities, such as atoms and bonds, followed by a compound structure reconstruction step. Recently, neural architectures analog to image captioning have been explored to solve this task, yet they still show to be data inefficient, using millions of examples just to show performances comparable with traditional methods. Looking to motivate and benchmark new approaches based on atomic-level entities detection and graph reconstruction, we present CEDe, a unique collection of chemical entity bounding boxes manually curated by experts for scientific literature datasets. These annotations combine to more than 700,000 chemical entity bounding boxes with the necessary information for structure reconstruction. Also, a large synthetic dataset containing one million molecular images and annotations is released in order to explore transfer-learning techniques that could help these architectures perform better under low-data regimes. Benchmarks show that detection-reconstruction based models can achieve performances on par with or better than image captioning-like models, even with 100x fewer training examples.

## Finding Naturally Occurring Physical Backdoors in Image Datasets

- Emily Wenger · Roma Bhattacharjee · Arjun Nitin Bhagoji · Josephine Passananti · Emilio Andere · Heather Zheng · Ben Zhao
- abstract@[open-review](#): Extensive literature on backdoor poison attacks has studied attacks and defenses for backdoors using digital trigger patterns. In contrast, physical backdoors use physical objects as triggers, have only recently been identified, and are qualitatively different enough to resist most defenses targeting digital trigger backdoors. Research on physical backdoors is limited by access to large datasets containing real images of physical objects co-located with misclassification targets. Building these datasets is time- and labor-intensive. This work seeks to address the challenge of accessibility for research on physical backdoor attacks. We hypothesize that there may be naturally occurring physically co-located objects already present in popular datasets such as ImageNet. Once identified, a careful relabeling of these data can transform them into training samples for physical backdoor attacks. We propose a method to scalably identify these subsets of potential triggers in existing datasets, along with the specific classes

they can poison. We call these naturally occurring trigger-class subsets natural backdoor datasets. Our techniques successfully identify natural backdoors in widely-available datasets, and produce models behaviorally equivalent to those trained on manually curated datasets. We release our code to allow the research community to create their own datasets for research on physical backdoor attacks.

## [ActionNet: A Multimodal Dataset for Human Activities Using Wearable Sensors in a Kitchen Environment](#)

- Joseph DelPreto · Chao Liu · Yiyue Luo · Michael Foshey · Yunzhu Li · Antonio Torralba · Wojciech Matusik · Daniela Rus
- abstract@[open-review](#): This paper introduces ActionNet, a multimodal dataset and recording framework with an emphasis on wearable sensing in a kitchen environment. It provides rich, synchronized data streams along with ground truth data to facilitate learning pipelines that could extract insights about how humans interact with the physical world during activities of daily living, and help lead to more capable and collaborative robot assistants. The wearable sensing suite captures motion, force, and attention information; it includes eye tracking with a first-person camera, forearm muscle activity sensors, a body-tracking system using 17 inertial sensors, finger-tracking gloves, and custom tactile sensors on the hands that use a matrix of conductive threads. This is coupled with activity labels and with externally-captured data from multiple RGB cameras, a depth camera, and microphones. The specific tasks recorded in ActionNet are designed to highlight lower-level physical skills and higher-level scene reasoning or action planning. They include simple object manipulations (e.g., stacking plates), dexterous actions (e.g., peeling or cutting vegetables), and complex action sequences (e.g., setting a table or loading a dishwasher). The resulting dataset and underlying experiment framework are available at <https://action-net.csail.mit.edu>. Preliminary networks and analyses explore modality subsets and cross-modal correlations. ActionNet aims to support applications including learning from demonstrations, dexterous robot control, cross-modal predictions, and fine-grained action segmentation. It could also help inform the next generation of smart textiles that may one day unobtrusively send rich data streams to in-home collaborative or autonomous robot assistants.

## [SkinCon: A skin disease dataset densely annotated by domain experts for fine-grained debugging and analysis](#)

- Roxana Daneshjou · Mert Yuksekgonul · Zhuso Ran Cai · Roberto Novoa · James Zou
- abstract@[open-review](#): For the deployment of artificial intelligence (AI) in high risk settings, such as healthcare, methods that provide interpretability/explainability or allow fine-grained error analysis are critical. Many recent methods for interpretability/explainability and fine-grained error analysis use concepts, which are meta-labels which are semantically meaningful to humans. However, there are only a few datasets that include concept-level meta-labels and most of these meta-labels are relevant for natural images that do not require domain expertise. Previous densely annotated datasets in medicine focused on meta-labels that are relevant to a single disease such as osteoarthritis or melanoma. In dermatology, skin disease is described using an established clinical lexicon that allow clinicians to describe physical exam findings to one another. To provide the first medical dataset densely annotated by domain experts to provide annotations useful across multiple disease processes, we developed SkinCon: a skin disease dataset densely annotated by dermatologists. SkinCon includes 3230 images from the Fitzpatrick 17k skin disease dataset densely annotated with 48 clinical concepts, 22 of which have at least 50 images representing the concept. The concepts used were chosen by two dermatologists considering the clinical descriptor terms used to describe skin lesions. Examples include "plaque", "scale", and "erosion". These same concepts were also used to label 656 skin disease images from the Diverse Dermatology Images dataset, providing an additional external dataset with diverse skin tone representations. We review the potential applications for the SkinCon dataset, such as probing models, concept-based explanations, concept bottlenecks, error analysis, and slice discovery. Furthermore, we use SkinCon to demonstrate two of these use cases: debugging mistakes of an existing dermatology AI model with concepts and developing interpretable models with post-hoc concept bottleneck models.

## [Myriad: a real-world testbed to bridge trajectory optimization and deep learning](#)

- Nikolaus Howe · Simon Dufort-Labbé · Nitarshan Rajkumar · Pierre-Luc Bacon
- abstract@[open-review](#): We present Myriad, a testbed written in JAX which enables machine learning researchers to benchmark imitation learning and reinforcement learning algorithms against trajectory optimization-based methods in real-world environments. Myriad contains 17 optimal control problems presented in continuous time which span medicine, ecology, epidemiology, and engineering. As such, Myriad strives to serve as a stepping stone towards application of modern machine learning techniques for impactful real-world tasks. The repository also provides machine learning practitioners access to trajectory optimization techniques, not only for standalone use, but also for integration within a typical automatic differentiation workflow. Indeed, the combination of classical control theory and deep learning in a fully GPU-compatible package unlocks potential for new algorithms to arise. We present one such novel approach for use in dynamics learning and control tasks. Trained in a fully end-to-end fashion, our model leverages an implicit planning module over neural ordinary differential equations, enabling simultaneous learning and planning with unknown environment dynamics. All environments, optimizers and tools are available in the software package at \url{https://github.com/nikihowe/myriad}.

## [MVP-N: A Dataset and Benchmark for Real-World Multi-View Object Classification](#)

- REN WANG · Jiayue Wang · Tae Sung Kim · JINSUNG KIM · Hyuk-Jae Lee
- abstract@[open-review](#): Combining information from multiple views is essential for discriminating similar objects. However, existing datasets for multi-view object classification have several limitations, such as synthetic and coarse-grained objects, no validation split for hyperparameter tuning, and a lack of view-level information quantity annotations for analyzing multi-view-based methods. To address this issue, this study proposes a new dataset, MVP-N, which contains 44 retail products, 16k real captured views with human-perceived information quantity annotations, and 9k multi-view sets. The fine-grained categorization of objects naturally generates multi-view label noise owing to the inter-class view similarity, allowing the study of learning from noisy labels in the multi-view case. Moreover, this study benchmarks four multi-view-based feature aggregation methods and twelve soft label methods on MVP-N. Experimental results show that MVP-N will be a valuable resource for facilitating the development of real-world multi-view object classification methods. The dataset and code are publicly available at <https://github.com/SMNUResearch/MVP-N>.

## [JAHS-Bench-201: A Foundation For Research On Joint Architecture And Hyperparameter Search](#)

- Archit Bansal · Danny Stoll · Maciej Janowski · Arber Zela · Frank Hutter
- abstract@[open-review](#): The past few years have seen the development of many benchmarks for Neural Architecture Search (NAS), fueling rapid progress in NAS research. However, recent work, that shows good hyperparameter settings can be more important than using the best architecture, calls for a shift in focus towards Joint Architecture and Hyperparameter Search (JAHS). Therefore, we present JAHS-Bench-201, the first collection of surrogate benchmarks for JAHS, built to also facilitate research on multi-objective, cost-aware and (multi) multi-fidelity optimization algorithms. To the best of our knowledge, JAHS-Bench-201 is based on the most extensive dataset of neural network performance data in the public domain. It is composed of approximately 140 million data points and 20 performance metrics for three deep learning tasks, while featuring a 14-dimensional search and fidelity space that extends the popular NAS-Bench-201 space. With JAHS-Bench-201, we hope to democratize research on JAHS and lower the barrier to entry of an extremely compute intensive field, e.g., by reducing the compute time to run a JAHS algorithm from 5 days to only a few seconds.

## [A Multi-Task Benchmark for Korean Legal Language Understanding and Judgement Prediction](#)

- Wonseok Hwang · Dongjun Lee · Kyoungyeon Cho · Hanuh Lee · Minjoon Seo
- abstract@[open-review](#): The recent advances of deep learning have dramatically changed how machine learning, especially in the domain of natural language processing, can be applied to legal domain. However, this shift to the data-driven approaches calls for larger and more diverse datasets, which are nevertheless still small in number, especially in non-English languages. Here we present the first large-scale benchmark of Korean legal AI datasets, LBOX OPEN, that consists of one legal corpus, two classification tasks, two legal judgement prediction (LJP) tasks, and one summarization task. The

legal corpus consists of 147k Korean precedents (259M tokens), of which 63k are sentenced in last 4 years and 96k are from the first and the second level courts in which factual issues are reviewed. The two classification tasks are case names (11.3k) and statutes (2.8k) prediction from the factual description of individual cases. The LJP tasks consist of (1) 10.5k criminal examples where the model is asked to predict fine amount, imprisonment with labor, and imprisonment without labor ranges for the given facts, and (2) 4.7k civil examples where the inputs are facts and claim for relief and outputs are the degrees of claim acceptance. The summarization task consists of the Supreme Court precedents and the corresponding summaries (20k). We also release realistic variants of the datasets by extending the domain (1) to infrequent case categories in case name (31k examples) and statute (17.7k) classification tasks, and (2) to long input sequences in the summarization task (51k). Finally, we release LCUBE, the first Korean legal language model trained on the legal corpus from this study. Given the uniqueness of the Law of South Korea and the diversity of the legal tasks covered in this work, we believe that LBOX OPEN contributes to the multilinguality of global legal research. LBOX OPEN and LCUBE will be publicly available.

## [Kantorovich Strikes Back! Wasserstein GANs are not Optimal Transport?](#)

- Alexander Korotin · Alexander Kolesov · Evgeny Burnaev
- abstract@[open-review](#): Wasserstein Generative Adversarial Networks (WGANs) are the popular generative models built on the theory of Optimal Transport (OT) and the Kantorovich duality. Despite the success of WGANs, it is still unclear how well the underlying OT dual solvers approximate the OT cost (Wasserstein-1 distance,  $W_1$ ) and the OT gradient needed to update the generator. In this paper, we address these questions. We construct 1-Lipschitz functions and use them to build ray monotone transport plans. This strategy yields pairs of continuous benchmark distributions with the analytically known OT plan, OT cost and OT gradient in high-dimensional spaces such as spaces of images. We thoroughly evaluate popular WGAN dual form solvers (gradient penalty, spectral normalization, entropic regularization, etc.) using these benchmark pairs. Even though these solvers perform well in WGANs, none of them faithfully compute  $W_1$  in high dimensions. Nevertheless, many provide a meaningful approximation of the OT gradient. These observations suggest that these solvers should not be treated as good estimators of  $W_1$  but to some extent they indeed can be used in variational problems requiring the minimization of  $W_1$ .

## [PEER: A Comprehensive and Multi-Task Benchmark for Protein Sequence Understanding](#)

- Minghao Xu · Zuobai Zhang · Jiarui Lu · Zhaocheng Zhu · Yangtian Zhang · Ma Chang · Runcheng Liu · Jian Tang
- abstract@[open-review](#): We are now witnessing significant progress of deep learning methods in a variety of tasks (or datasets) of proteins. However, there is a lack of a standard benchmark to evaluate the performance of different methods, which hinders the progress of deep learning in this field. In this paper, we propose such a benchmark called PEER, a comprehensive and multi-task benchmark for Protein sEquence undERstanding. PEER provides a set of diverse protein understanding tasks including protein function prediction, protein localization prediction, protein structure prediction, protein-protein interaction prediction, and protein-ligand interaction prediction. We evaluate different types of sequence-based methods for each task including traditional feature engineering approaches, different sequence encoding methods as well as large-scale pre-trained protein language models. In addition, we also investigate the performance of these methods under the multi-task learning setting. Experimental results show that large-scale pre-trained protein language models achieve the best performance for most individual tasks, and jointly training multiple tasks further boosts the performance. The datasets and source codes of this benchmark will be open-sourced soon.

## [A Large Scale Search Dataset for Unbiased Learning to Rank](#)

- Lixin Zou · Haitao Mao · Xiaokai Chu · Jiliang Tang · Wenwen Ye · Shuaiqiang Wang · Dawei Yin
- abstract@[open-review](#): The unbiased learning to rank (ULTR) problem has been greatly advanced by recent deep learning techniques and well-designed debias algorithms. However, promising results on the existing benchmark datasets may not be extended to the practical scenario due to some limitations of existing datasets. First, their semantic feature extractions are outdated while state-of-the-art large-scale pre-trained language models like BERT cannot be utilized due to the lack of original text. Second, display features are incomplete; thus in-depth study on ULTR is impossible such as the displayed abstract for analyzing the click necessary bias. Third, synthetic user feedback has been adopted by most existing datasets and real-world user feedback is greatly missing. To overcome these disadvantages, we introduce the Baidu-ULTR dataset. It involves randomly sampled 1.2 billion searching sessions and 7,008 expert annotated queries(397,572 query document pairs). Baidu-ULTR is the first billion-level dataset for ULTR. Particularly, it offers: (1)the original semantic features and pre-trained language models of different sizes; (2)sufficient display information such as position, displayed height, and displayed abstract, enabling the comprehensive study of multiple displayed biases; and (3)rich user feedback on search result pages (SERPs) like dwelling time, allowing for user engagement optimization and promoting the exploration of multi-task learning in ULTR. Furthermore, we present the design principle of Baidu-ULTR and the performance of representative ULTR algorithms on Baidu-ULTR. The Baidu-ULTR dataset and corresponding baseline implementations are available at <https://github.com/ChuXiaokai/baiduultrdataset>. The dataset homepage is available at <https://searchscience.baidu.com/dataset.html>.

## [AMOS: A Large-Scale Abdominal Multi-Organ Benchmark for Versatile Medical Image Segmentation](#)

- Yuanfeng Ji · Haotian Bai · Chongjian GE · Jie Yang · Ye Zhu · Ruimao Zhang · Zhen Li · Lingyan Zhanng · Wanling Ma · Xiang Wan · Ping Luo
- abstract@[open-review](#): Despite the considerable progress in automatic abdominal multi-organ segmentation from CT/MRI scans in recent years, a comprehensive evaluation of the models' capabilities is hampered by the lack of a large-scale benchmark from diverse clinical scenarios. Constraint by the high cost of collecting and labeling 3D medical data, most of the deep learning models to date are driven by datasets with a limited number of organs of interest or samples, which still limits the power of modern deep models and makes it difficult to provide a fully comprehensive and fair estimate of various methods. To mitigate the limitations, we present AMOS, a large-scale, diverse, clinical dataset for abdominal organ segmentation. AMOS provides 500 CT and 100 MRI scans collected from multi-center, multi-vendor, multi-modality, multi-phase, multi-disease patients, each with voxel-level annotations of 15 abdominal organs, providing challenging examples and test-bed for studying robust segmentation algorithms under diverse targets and scenarios. We further benchmark several state-of-the-art medical segmentation models to evaluate the status of the existing methods on this new challenging dataset. We have made our datasets, benchmark servers, and baselines publicly available, and hope to inspire future research. Information can be found at <https://amos22.grand-challenge.org>.

## [PeRFception: Perception using Radiance Fields](#)

- Yoonwoo Jeong · Seungjoo Shin · Junha Lee · Chris Choy · Anima Anandkumar · Minsu Cho · Jaesik Park
- abstract@[open-review](#): The recent progress in implicit 3D representation, i.e., Neural Radiance Fields (NeRFs), has made accurate and photorealistic 3D reconstruction possible in a differentiable manner. This new representation can effectively convey the information of hundreds of high-resolution images in one compact format and allows photorealistic synthesis of novel views. In this work, using the variant of NeRF called Plenoxels, we create the first large-scale radiance fields datasets for perception tasks, called the PeRFception, which consists of two parts that incorporate both object-centric and scene-centric scans for classification and segmentation. It shows a significant memory compression rate (96.4%) from the original dataset, while containing both 2D and 3D information in a unified form. We construct the classification and segmentation models that directly take this radiance fields format as input and also propose a novel augmentation technique to avoid overfitting on backgrounds of images. The code and data are publicly available in "<https://postech-cvlab.github.io/PeRFception/>".

## [BLOX: Macro Neural Architecture Search Benchmark and Algorithms](#)

- Thomas Chau Å· Łukasz Dudziak Å· Hongkai Wen Å· Nicholas Lane Å· Mohamed Abdelfattah
- abstract@[open-review](#): Neural architecture search (NAS) has been successfully used to design numerous high-performance neural networks. However, NAS is typically compute-intensive, so most existing approaches restrict the search to decide the operations and topological structure of a single block only, then the same block is stacked repeatedly to form an end-to-end model. Although such an approach reduces the size of search space, recent studies show that a macro search space, which allows blocks in a model to be different, can lead to better performance. To provide a systematic study of the performance of NAS algorithms on a macro search space, we release Blox — a benchmark that consists of 91k unique models trained on the CIFAR-100 dataset. The dataset also includes runtime measurements of all the models on a diverse set of hardware platforms. We perform extensive experiments to compare existing algorithms that are well studied on cell-based search spaces, with the emerging blockwise approaches that aim to make NAS scalable to much larger macro search spaces. The Blox benchmark and code are available at <https://github.com/SamsungLabs/blox>.

## NAS-Bench-Graph: Benchmarking Graph Neural Architecture Search

- Yijian Qin Å· Ziwei Zhang Å· Xin Wang Å· Zeyang Zhang Å· Wenwu Zhu
- abstract@[open-review](#): Graph neural architecture search (GraphNAS) has recently aroused considerable attention in both academia and industry. However, two key challenges seriously hinder the further research of GraphNAS. First, since there is no consensus for the experimental setting, the empirical results in different research papers are often not comparable and even not reproducible, leading to unfair comparisons. Secondly, GraphNAS often needs extensive computations, which makes it highly inefficient and inaccessible to researchers without access to large-scale computation. To solve these challenges, we propose NAS-Bench-Graph, a tailored benchmark that supports unified, reproducible, and efficient evaluations for GraphNAS. Specifically, we construct a unified, expressive yet compact search space, covering 26,206 unique graph neural network (GNN) architectures and propose a principled evaluation protocol. To avoid unnecessary repetitive training, we have trained and evaluated all of these architectures on nine representative graph datasets, recording detailed metrics including train, validation, and test performance in each epoch, the latency, the number of parameters, etc. Based on our proposed benchmark, the performance of GNN architectures can be directly obtained by a look-up table without any further computation, which enables fair, fully reproducible, and efficient comparisons. To demonstrate its usage, we make in-depth analyses of our proposed NAS-Bench-Graph, revealing several interesting findings for GraphNAS. We also showcase how the benchmark can be easily compatible with GraphNAS open libraries such as AutoGL and NNI. To the best of our knowledge, our work is the first benchmark for graph neural architecture search.

## Sample Efficiency Matters: A Benchmark for Practical Molecular Optimization

- Wenhao Gao Å· Tianfan Fu Å· Jimeng Sun Å· Connor Coley
- abstract@[open-review](#): Molecular optimization is a fundamental goal in the chemical sciences and is of central interest to drug and material design. In recent years, significant progress has been made in solving challenging problems across various aspects of computational molecular optimizations, emphasizing high validity, diversity, and, most recently, synthesizability. Despite this progress, many papers report results on trivial or self-designed tasks, bringing additional challenges to directly assessing the performance of new methods. Moreover, the sample efficiency of the optimization—the number of molecules evaluated by the oracle—is rarely discussed, despite being an essential consideration for realistic discovery applications. To fill this gap, we have created an open-source benchmark for practical molecular optimization, PMO, to facilitate the transparent and reproducible evaluation of algorithmic advances in molecular optimization. This paper thoroughly investigates the performance of 25 molecular design algorithms on 23 single-objective (scalar) optimization tasks with a particular focus on sample efficiency. Our results show that most “state-of-the-art” methods fail to outperform their predecessors under a limited oracle budget allowing 10K queries and that no existing algorithm can efficiently solve certain molecular optimization problems in this setting. We analyze the influence of the optimization algorithm choices, molecular assembly strategies, and oracle landscapes on the optimization performance to inform future algorithm development and benchmarking. PMO provides a standardized experimental setup to comprehensively evaluate and compare new molecule optimization methods with existing ones. All code can be found at [https://github.com/wenhao-gao/mol\\_opt](https://github.com/wenhao-gao/mol_opt).

## This is the way: designing and compiling LEPISZCZE, a comprehensive NLP benchmark for Polish

- Łukasz Augustyniak Å· Kamil Tagowski Å· Albert Sawczyn Å· Denis Janiak Å· Roman Bartusiak Å· Adrian Szymczak Å· Arkadiusz Janz Å· Piotr SzymaÅ„ski Å· Marcin WÄ...roba Å· MikoÅ„aj Morzy Å· Tomasz Kajdanowicz Å· Maciej Piasecki
- abstract@[open-review](#): The availability of compute and data to train larger and larger language models increases the demand for robust methods of benchmarking the true progress of LM training. Recent years witnessed significant progress in standardized benchmarking for English. Benchmarks such as GLUE, SuperGLUE, or KILT have become a de facto standard tools to compare large language models. Following the trend to replicate GLUE for other languages, the KLEJ benchmark (klej is the word for glue in Polish) has been released for Polish. In this paper, we evaluate the progress in benchmarking for low-resourced languages. We note that only a handful of languages have such comprehensive benchmarks. We also note the gap in the number of tasks being evaluated by benchmarks for resource-rich English/Chinese and the rest of the world. In this paper, we introduce LEPISZCZE (lepiszcze is the Polish word for glew, the Middle English predecessor of glue), a new, comprehensive benchmark for Polish NLP with a large variety of tasks and high-quality operationalization of the benchmark. We design LEPISZCZE with flexibility in mind. Including new models, datasets, and tasks is as simple as possible while still offering data versioning and model tracking. In the first run of the benchmark, we test 13 experiments (task and dataset pairs) based on the five most recent LMs for Polish. We use five datasets from the Polish benchmark and add eight novel datasets. As the paper’s main contribution, apart from LEPISZCZE, we provide insights and experiences learned while creating the benchmark for Polish as the blueprint to design similar benchmarks for other low-resourced languages.

## Benchmarking and Analyzing 3D Human Pose and Shape Estimation Beyond Algorithms

- Hui En Pang Å· Zhongang Cai Å· Lei Yang Å· Tianwei Zhang Å· Ziwei Liu
- abstract@[open-review](#): 3D human pose and shape estimation (a.k.a. “human mesh recovery”) has achieved substantial progress. Researchers mainly focus on the development of novel algorithms, while less attention has been paid to other critical factors involved. This could lead to less optimal baselines, hindering the fair and faithful evaluations of newly designed methodologies. To address this problem, this work presents the first comprehensive benchmarking study from three under-explored perspectives beyond algorithms. 1) Datasets. An analysis on 31 datasets reveals the distinct impacts of data samples: datasets featuring critical attributes (i.e., diverse poses, shapes, camera characteristics, backbone features) are more effective. Strategical selection and combination of high-quality datasets can yield a significant boost to the model performance. 2) Backbones. Experiments with 10 backbones, ranging from CNNs to transformers, show the knowledge learnt from a proximity task is readily transferable to human mesh recovery. 3) Training strategies. Proper augmentation techniques and loss designs are crucial. With the above findings, we achieve a PA-MPJPE of 47.3 (mm) on the 3DPW test set with a relatively simple model. More importantly, we provide strong baselines for fair comparisons of algorithms, and recommendations for building effective training configurations in the future. Codebase is available at <https://github.com/smplbody/hmr-benchmarks>.

## MATE: Benchmarking Multi-Agent Reinforcement Learning in Distributed Target Coverage Control

- Xuehai Pan Å· Mickel Liu Å· Fangwei Zhong Å· Yaodong Yang Å· Song-Chun Zhu Å· Yizhou Wang
- abstract@[open-review](#): We introduce the Multi-Agent Tracking Environment (MATE), a novel multi-agent environment simulates the target coverage control problems in the real world. MATE hosts an asymmetric cooperative-competitive game consisting of two groups of learning agents—“cameras” and “targets”—with opposing interests. Specifically, “cameras”, a group of directional sensors, are mandated to actively control the directional perception area to maximize the coverage rate of targets. On the other side, “targets” are mobile agents that aim to transport cargo between multiple randomly assigned warehouses while minimizing the exposure to the camera sensor networks. To showcase the practicality of MATE, we benchmark the multi-agent

reinforcement learning (MARL) algorithms from different aspects, including cooperation, communication, scalability, robustness, and asymmetric self-play. We start by reporting results for cooperative tasks using MARL algorithms (MAPPO, IPPO, QMIX, MADDPG) and the results after augmenting with multi-agent communication protocols (TarMAC, I2C). We then evaluate the effectiveness of the popular self-play techniques (PSRO, fictitious self-play) in an asymmetric zero-sum competitive game. This process of co-evolution between cameras and targets helps to realize a less exploitable camera network. We also observe the emergence of different roles of the target agents while incorporating I2C into target-target communication. MATE is written purely in Python and integrated with OpenAI Gym API to enhance user-friendliness. Our project is released at <https://github.com/UnrealTracking/mate>.

## [METS-CoV: A Dataset of Medical Entity and Targeted Sentiment on COVID-19 Related Tweets](#)

- Peilin Zhou · Zeqiang Wang · Dading Chong · Zhijiang Guo · Yining Hua · Zichang Su · Zhiyang Teng · Jiageng Wu · Jie Yang
- abstract@[open-review](#): The COVID-19 pandemic continues to bring up various topics discussed or debated on social media. In order to explore the impact of pandemics on people's lives, it is crucial to understand the public's concerns and attitudes towards pandemic-related entities (e.g., drugs, vaccines) on social media. However, models trained on existing named entity recognition (NER) or targeted sentiment analysis (TSA) datasets have limited ability to understand COVID-19-related social media texts because these datasets are not designed or annotated from a medical perspective. In this paper, we release METS-CoV, a dataset containing medical entities and targeted sentiments from COVID-19 related tweets. METS-CoV contains 10,000 tweets with 7 types of entities, including 4 medical entity types (Disease, Drug, Symptom, and Vaccine) and 3 general entity types (Person, Location, and Organization). To further investigate tweet users' attitudes toward specific entities, 4 types of entities (Person, Organization, Drug, and Vaccine) are selected and annotated with user sentiments, resulting in a targeted sentiment dataset with 9,101 entities (in 5,278 tweets). To the best of our knowledge, METS-CoV is the first dataset to collect medical entities and corresponding sentiments of COVID-19 related tweets. We benchmark the performance of classical machine learning models and state-of-the-art deep learning models on NER and TSA tasks with extensive experiments. Results show that this dataset has vast room for improvement for both NER and TSA tasks. With rich annotations and comprehensive benchmark results, we believe METS-CoV is a fundamental resource for building better medical social media understanding tools and facilitating computational social science research, especially on epidemiological topics. Our data, annotation guidelines, benchmark models, and source code are publicly available (<https://github.com/YLab-Open/METS-CoV>) to ensure reproducibility.

## [Geoclideo: Few-Shot Generalization in Euclidean Geometry](#)

- Joy Hsu · Jiajun Wu · Noah Goodman
- abstract@[open-review](#): Euclidean geometry is among the earliest forms of mathematical thinking. While the geometric primitives underlying its constructions, such as perfect lines and circles, do not often occur in the natural world, humans rarely struggle to perceive and reason with them. Will computer vision models trained on natural images show the same sensitivity to Euclidean geometry? Here we explore these questions by studying few-shot generalization in the universe of Euclidean geometry constructions. We introduce Geoclideo, a domain-specific language for Euclidean geometry, and use it to generate two datasets of geometric concept learning tasks for benchmarking generalization judgements of humans and machines. We find that humans are indeed sensitive to Euclidean geometry and generalize strongly from a few visual examples of a geometric concept. In contrast, low-level and high-level visual features from standard computer vision models pretrained on natural images do not support correct generalization. Thus Geoclideo represents a novel few-shot generalization benchmark for geometric concept learning, where the performance of humans and of AI models diverge. The Geoclideo framework and dataset are publicly available for download.

## [How Transferable are Video Representations Based on Synthetic Data?](#)

- Yo-whan Kim · Samarth Mishra · SouYoung Jin · Rameswar Panda · Hilde Kuehne · Leonid Karlinsky · Venkatesh Saligrama · Kate Saenko · Aude Oliva · Rogerio Feris
- abstract@[open-review](#): Action recognition has improved dramatically with massive-scale video datasets. Yet, these datasets are accompanied with issues related to curation cost, privacy, ethics, bias, and copyright. Compared to that, only minor efforts have been devoted toward exploring the potential of synthetic video data. In this work, as a stepping stone towards addressing these shortcomings, we study the transferability of video representations learned solely from synthetically-generated video clips, instead of real data. We propose SynAPT, a novel benchmark for action recognition based on a combination of existing synthetic datasets, in which a model is pre-trained on synthetic videos rendered by various graphics simulators, and then transferred to a set of downstream action recognition datasets, containing different categories than the synthetic data. We provide an extensive baseline analysis on SynAPT revealing that the simulation-to-real gap is minor for datasets with low object and scene bias, where models pre-trained with synthetic data even outperform their real data counterparts. We posit that the gap between real and synthetic action representations can be attributed to contextual bias and static objects related to the action, instead of the temporal dynamics of the action itself. The SynAPT benchmark is available at <https://github.com/mintjohnkim/SynAPT>.

## [Open High-Resolution Satellite Imagery: The WorldStrat Dataset â€“ With Application to Super-Resolution](#)

- Julien Cornebise · Ivan Orjoli · Freddie Kalaitzis
- abstract@[open-review](#): Analyzing the planet at scale with satellite imagery and machine learning is a dream that has been constantly hindered by the cost of difficult-to-access highly-representative high-resolution imagery. To remediate this, we introduce here the WorldStratified dataset. The largest and most varied such publicly available dataset, at Airbus SPOT 6/7 satellites' high resolution of up to 1.5 m/pixel, empowered by European Space Agency's Phi-Lab as part of the ESA-funded QueryPlanet project, we curate 10,000 sq km of unique locations to ensure stratified representation of all types of land-use across the world: from agriculture to ice caps, from forests to multiple urbanization densities. We also enrich those with locations typically under-represented in ML datasets: sites of humanitarian interest, illegal mining sites, and settlements of persons at risk. We temporally-match each high-resolution image with multiple low-resolution images from the freely accessible lower-resolution Sentinel-2 satellites at 10 m/pixel. We accompany this dataset with an open-source Python package to: rebuild or extend the WorldStrat dataset, train and infer baseline algorithms, and learn with abundant tutorials, all compatible with the popular EO-learn toolbox. We hereby hope to foster broad-spectrum applications of ML to satellite imagery, and possibly develop from free public low-resolution Sentinel2 imagery the same power of analysis allowed by costly private high-resolution imagery. We illustrate this specific point by training and releasing several highly compute-efficient baselines on the task of Multi-Frame Super-Resolution. License-wise, the high-resolution Airbus imagery is CC-BY-NC, while the labels, Sentinel2 imagery, and trained weights are under CC-BY, and the source code under BSD, to allow for the widest use and dissemination. The dataset is available at <https://zenodo.org/record/6810792> and the software package at <https://github.com/worldstrat/worldstrat>.

## [Wild-Time: A Benchmark of in-the-Wild Distribution Shift over Time](#)

- Huaxiu Yao · Caroline Choi · Bochuan Cao · Yoonho Lee · Pang Wei Koh · Chelsea Finn
- abstract@[open-review](#): Distribution shifts occur when the test distribution differs from the training distribution, and can considerably degrade performance of machine learning models deployed in the real world. While recent works have studied robustness to distribution shifts, distribution shifts arising from the passage of time have the additional structure of timestamp metadata. Real-world examples of such shifts are underexplored, and it is unclear whether existing models can leverage trends in past distribution shifts to reliably extrapolate into the future. To address this gap, we curate Wild-Time, a benchmark of 7 datasets that reflect temporal distribution shifts arising in a variety of real-world applications, including drug discovery, patient prognosis, and news classification. On these datasets, we systematically benchmark 13 approaches with various inductive biases. We evaluate methods in domain-generalization, continual learning, self-supervised learning, and ensemble learning, which leverage timestamps to extract the common structure of the distribution shifts. We extend several domain-generalization methods to the temporal distribution shift setting by treating windows of time as different

domains. Finally, we propose two evaluation strategies to evaluate model performance under temporal distribution shifts---evaluation with a fixed time split (Eval-Fix) and evaluation with a data stream (Eval-Stream). Eval-Fix, our primary evaluation strategy, aims to provide a simple evaluation protocol for the broader machine learning community, while Eval-Stream serves as a complementary benchmark for continual learning approaches. Our experiments demonstrate that existing methods are limited in tackling temporal distribution shift: across all settings, we observe an average performance drop of 20% from in-distribution to out-of-distribution data.

## [Pile of Law: Learning Responsible Data Filtering from the Law and a 256GB Open-Source Legal Dataset](#)

- Peter Henderson · Mark Krass · Lucia Zheng · Neel Guha · Christopher D Manning · Dan Jurafsky · Daniel Ho
- abstract@[open-review](#): One concern with the rise of large language models lies with their potential for significant harm, particularly from pretraining on biased, obscene, copyrighted, and private information. Emerging ethical approaches have attempted to filter pretraining material, but such approaches have been ad hoc and failed to take context into account. We offer an approach to filtering grounded in law, which has directly addressed the tradeoffs in filtering material. First, we gather and make available the Pile of Law, a ~256GB (and growing) dataset of open-source English-language legal and administrative data, covering court opinions, contracts, administrative rules, and legislative records. Pretraining on the Pile of Law may help with legal tasks that have the promise to improve access to justice. Second, we distill the legal norms that governments have developed to constrain the inclusion of toxic or private content into actionable lessons for researchers and discuss how our dataset reflects these norms. Third, we show how the Pile of Law offers researchers the opportunity to learn such filtering rules directly from the data, providing an exciting new research direction in model-based processing.

## [AirfRANS: High Fidelity Computational Fluid Dynamics Dataset for Approximating Reynolds-Averaged Navierâ€“Stokes Solutions](#)

- Florent Bonnet · Jocelyn Mazari · Paola Cinnella · Patrick Gallinari
- abstract@[open-review](#): Surrogate models are necessary to optimize meaningful quantities in physical dynamics as their recursive numerical resolutions are often prohibitively expensive. It is mainly the case for fluid dynamics and the resolution of Navierâ€“Stokes equations. However, despite the fast-growing field of data-driven models for physical systems, reference datasets representing real-world phenomena are lacking. In this work, we develop \textsc{AirfRANS}, a dataset for studying the two-dimensional incompressible steady-state Reynolds-Averaged Navierâ€“Stokes equations over airfoils at a subsonic regime and for different angles of attacks. We also introduce metrics on the stress forces at the surface of geometries and visualization of boundary layers to assess the capabilities of models to accurately predict the meaningful information of the problem. Finally, we propose deep learning baselines on four machine learning tasks to study \textsc{AirfRANS} under different constraints for generalization considerations: big and scarce data regime, Reynolds number, and angle of attack extrapolation.

## [Towards Human-Level Bimanual Dexterous Manipulation with Reinforcement Learning](#)

- Yuanpei Chen · Tianhao Wu · Shengjie Wang · Xidong Feng · Jiechuan Jiang · Zongqing Lu · Stephen McAleer · Hao Dong · Song-Chun Zhu · Yaodong Yang
- abstract@[open-review](#): Achieving human-level dexterity is an important open problem in robotics. However, tasks of dexterous hand manipulation even at the baby level are challenging to solve through reinforcement learning (RL). The difficulty lies in the high degrees of freedom and the required cooperation among heterogeneous agents (e.g., joints of fingers). In this study, we propose the Bimanual Dexterous Hands Benchmark (Bi-DexHands), a simulator that involves two dexterous hands with tens of bimanual manipulation tasks and thousands of target objects. Tasks in Bi-DexHands are first designed to match human-level motor skills according to literature in cognitive science, and then are built in Isaac Gym; this enables highly efficient RL trainings, reaching 30,000+ FPS by only one single NVIDIA RTX 3090. We provide a comprehensive benchmark for popular RL algorithms under different settings; this includes multi-agent RL, offline RL, multi-task RL, and meta RL. Our results show that PPO type on-policy algorithms can learn to solve simple manipulation tasks that are equivalent up to 48-month human baby (e.g., catching a flying object, opening a bottle), while multi-agent RL can further help to learn manipulations that require skilled bimanual cooperation (e.g., lifting a pot, stacking blocks). Despite the success on each individual task, when it comes to mastering multiple manipulation skills, existing RL algorithms fail to work in most of the multi-task and the few-shot learning tasks, which calls for more future development from the RL community. Our project is open-sourced at <https://github.com/PKU-MARL/DexterousHands>.

## [LIPS - Learning Industrial Physical Simulation benchmark suite](#)

- Milad LEYLI ABADI · Antoine Marot · JÃ©rÃ©me Picault · David Danan · Mouadh Yagoubi · Benjamin Donnot · Seif Attoui · Pavel Dimitrov · Asma Farjallah · Clement Etienam
- abstract@[open-review](#): Physical simulations are at the core of many critical industrial systems. However, today's physical simulators have some limitations such as computation time, dealing with missing or uncertain data, or even non-convergence for some feasible cases. Recently, the use of data-driven approaches to learn complex physical simulations has been considered as a promising approach to address those issues. However, this comes often at the cost of some accuracy which may hinder the industrial use. To drive this new research topic towards a better real-world applicability, we propose a new benchmark suite "Learning Industrial Physical Simulations"(LIPS) to meet the need of developing efficient, industrial application-oriented, augmented simulators. To define how to assess such benchmark performance, we propose a set of four generic categories of criteria. The proposed benchmark suite is a modular and configurable framework that can deal with different physical problems. To demonstrate this ability, we propose in this paper to investigate two distinct use-cases with different physical simulations, namely: the power grid and the pneumatic. For each use case, several benchmarks are described and assessed with existing models. None of the models perform well under all expected criteria, inviting the community to develop new industry-applicable solutions and possibly showcase their performance publicly upon online LIPS instance on Codabench.

## [NAS-Bench-Suite-Zero: Accelerating Research on Zero Cost Proxies](#)

- Arjun Krishnakumar · Colin White · Arber Zela · Renbo Tu · Mahmoud Safari · Frank Hutter
- abstract@[open-review](#): Zero-cost proxies (ZC proxies) are a recent architecture performance prediction technique aiming to significantly speed up algorithms for neural architecture search (NAS). Recent work has shown that these techniques show great promise, but certain aspects, such as evaluating and exploiting their complementary strengths, are under-studied. In this work, we create NAS-Bench-Suite: we evaluate 13 ZC proxies across 28 tasks, creating by far the largest dataset (and unified codebase) for ZC proxies, enabling orders-of-magnitude faster experiments on ZC proxies, while avoiding confounding factors stemming from different implementations. To demonstrate the usefulness of NAS-Bench-Suite, we run a large-scale analysis of ZC proxies, including a bias analysis, and the first information-theoretic analysis which concludes that ZC proxies capture substantial complementary information. Motivated by these findings, we present a procedure to improve the performance of ZC proxies by reducing biases such as cell size, and we also show that incorporating all 13 ZC proxies into the surrogate models used by NAS algorithms can improve their predictive performance by up to 42%. Our code and datasets are available at <https://github.com/automl/naslib/tree/zerocost>.

## [EPIC-KITCHENS VISOR Benchmark: VVideo Segmentations and Object Relations](#)

- Ahmad Darkhalil · Dandan Shan · Bin Zhu · Jian Ma · Amlan Kar · Richard Higgins · Sanja Fidler · David Fouhey · Dima Damen
- abstract@[open-review](#): We introduce VISOR, a new dataset of pixel annotations and a benchmark suite for segmenting hands and active objects in egocentric video. VISOR annotates videos from EPIC-KITCHENS, which comes with a new set of challenges not encountered in current video segmentation datasets. Specifically, we need to ensure both short- and long-term consistency of pixel-level annotations as objects undergo transformative

interactions, e.g. an onion is peeled, diced and cooked - where we aim to obtain accurate pixel-level annotations of the peel, onion pieces, chopping board, knife, pan, as well as the acting hands. VISOR introduces an annotation pipeline, AI-powered in parts, for scalability and quality. In total, we publicly release 272K manual semantic masks of 257 object classes, 9.9M interpolated dense masks, 67K hand-object relations, covering 36 hours of 179 untrimmed videos. Along with the annotations, we introduce three challenges in video object segmentation, interaction understanding and long-term reasoning. For data, code and leaderboards: <http://epic-kitchens.github.io/VISOR>

## [ELEVATER: A Benchmark and Toolkit for Evaluating Language-Augmented Visual Models](#)

- Chunyuan Li · Haotian Liu · Liunian Li · Pengchuan Zhang · Jyoti Aneja · Jianwei Yang · Ping Jin · Houdong Hu · Zicheng Liu · Yong Jae Lee · Jianfeng Gao
- abstract@[open-review](#): Learning visual representations from natural language supervision has recently shown great promise in a number of pioneering works. In general, these language-augmented visual models demonstrate strong transferability to a variety of datasets/tasks. However, it remains challenging to evaluate the transferability of these foundation models due to the lack of easy-to-use toolkits for fair benchmarking. To tackle this, we build ELEVATER (Evaluation of Language-augmented Visual Task-level Transfer), the first benchmark to compare and evaluate pre-trained language-augmented visual models. Several highlights include: (i) Datasets. As downstream evaluation suites, it consists of 20 image classification datasets and 35 object detection datasets, each of which is augmented with external knowledge. (ii) Toolkit. An automatic hyper-parameter tuning toolkit is developed to ensure the fairness in model adaption. To leverage the full power of language-augmented visual models, novel language-aware initialization methods are proposed to significantly improve the adaption performance. (iii) Metrics. A variety of evaluation metrics are used, including sample-efficiency (zero-shot and few-shot) and parameter-efficiency (linear probing and full model fine-tuning). We will publicly release ELEVATER.

## [Touch and Go: Learning from Human-Collected Vision and Touch](#)

- Fengyu Yang · Chenyang Ma · Jiacheng Zhang · Jing Zhu · Wenzhen Yuan · Andrew Owens
- abstract@[open-review](#): The ability to associate sight with touch is essential for understanding material properties, and for physically interacting with the world. Learning these correlations, however, has proven challenging, since existing datasets have not captured the full diversity of these modalities. To address this shortcoming, we propose a dataset for multimodal visuo-tactile learning called Touch and Go, in which human data collectors probe objects in natural environments with tactile sensors, while recording egocentric video. The objects and scenes in our dataset are significantly more diverse than prior efforts, making the data well-suited to tasks that involve understanding material properties and physical interactions in the wild. To demonstrate our dataset's effectiveness, we successfully apply it to a variety of tasks: 1) self-supervised visuo-tactile feature learning, 2) tactile-driven image stylization, i.e., making the visual appearance of an object more consistent with a given tactile signal, and 3) predicting future frames of a tactile signal from visuo-tactile inputs.

## [MoCapAct: A Multi-Task Dataset for Simulated Humanoid Control](#)

- Nolan Wagener · Andrey Kolobov · Felipe Vieira Frujeri · Ricky Loynd · Ching-An Cheng · Matthew Hausknecht
- abstract@[open-review](#): Simulated humanoids are an appealing research domain due to their physical capabilities. Nonetheless, they are also challenging to control, as a policy must drive an unstable, discontinuous, and high-dimensional physical system. One widely studied approach is to utilize motion capture (MoCap) data to teach the humanoid agent low-level skills (e.g., standing, walking, and running) that can then be re-used to synthesize high-level behaviors. However, even with MoCap data, controlling simulated humanoids remains very hard, as MoCap data offers only kinematic information. Finding physical control inputs to realize the demonstrated motions requires computationally intensive methods like reinforcement learning. Thus, despite the publicly available MoCap data, its utility has been limited to institutions with large-scale compute. In this work, we dramatically lower the barrier for productive research on this topic by training and releasing high-quality agents that can track over three hours of MoCap data for a simulated humanoid in the dmcontrol physics-based environment. We release MoCapAct (Motion Capture with Actions), a dataset of these expert agents and their rollouts, which contain proprioceptive observations and actions. We demonstrate the utility of MoCapAct by using it to train a single hierarchical policy capable of tracking the entire MoCap dataset within dmcontrol and show the learned low-level component can be re-used to efficiently learn downstream high-level tasks. Finally, we use MoCapAct to train an autoregressive GPT model and show that it can control a simulated humanoid to perform natural motion completion given a motion prompt. Videos of the results and links to the code and dataset are available at <https://microsoft.github.io/MoCapAct>.

## [DDXPlus: A New Dataset For Automatic Medical Diagnosis](#)

- Arsene Fansi Tchang · Rishab Goel · Zhi Wen · Julien Martel · Joumana Ghosn
- abstract@[open-review](#): There has been a rapidly growing interest in Automatic Symptom Detection (ASD) and Automatic Diagnosis (AD) systems in the machine learning research literature, aiming to assist doctors in telemedicine services. These systems are designed to interact with patients, collect evidence about their symptoms and relevant antecedents, and possibly make predictions about the underlying diseases. Doctors would review the interactions, including the evidence and the predictions, collect if necessary additional information from patients, before deciding on next steps. Despite recent progress in this area, an important piece of doctors' interactions with patients is missing in the design of these systems, namely the differential diagnosis. Its absence is largely due to the lack of datasets that include such information for models to train on. In this work, we present a large-scale synthetic dataset of roughly 1.3 million patients that includes a differential diagnosis, along with the ground truth pathology, symptoms and antecedents for each patient. Unlike existing datasets which only contain binary symptoms and antecedents, this dataset also contains categorical and multi-choice symptoms and antecedents useful for efficient data collection. Moreover, some symptoms are organized in a hierarchy, making it possible to design systems able to interact with patients in a logical way. As a proof-of-concept, we extend two existing AD and ASD systems to incorporate the differential diagnosis, and provide empirical evidence that using differentials as training signals is essential for the efficiency of such systems or for helping doctors better understand the reasoning of those systems.

## [FACT: Learning Governing Abstractions Behind Integer Sequences](#)

- Peter Belcak · Ard Kastrati · Flavio Schenker · Roger Wattenhofer
- abstract@[open-review](#): Integer sequences are of central importance to the modeling of concepts admitting complete finitary descriptions. We introduce a novel view on the learning of such concepts and lay down a set of benchmarking tasks aimed at conceptual understanding by machine learning models. These tasks indirectly assess model ability to abstract, and challenge them to reason both interpolatively and extrapolatively from the knowledge gained by observing representative examples. To further aid research in knowledge representation and reasoning, we present FACT, the Finitary Abstraction Comprehension Toolkit. The toolkit surrounds a large dataset of integer sequences comprising both organic and synthetic entries, a library for data pre-processing and generation, a set of model performance evaluation tools, and a collection of baseline model implementations, enabling the making of the future advancements with ease.

## [AutoWS-Bench-101: Benchmarking Automated Weak Supervision with 100 Labels](#)

- Nicholas Roberts · Xintong Li · Tzu-Heng Huang · Dyah Adila · Spencer Schoenberg · Cheng-Yu Liu · Lauren Pick · Haotian Ma · Aws Albargouthi · Frederic Sala
- abstract@[open-review](#): Weak supervision (WS) is a powerful method to build labeled datasets for training supervised models in the face of little-to-no labeled data. It replaces hand-labeling data with aggregating multiple noisy-but-cheap label estimates expressed by labeling functions (LFs). While it has been used successfully in many domains, weak supervision's application scope is limited by the difficulty of constructing labeling functions for domains

with complex or high-dimensional features. To address this, a handful of methods have proposed automating the LF design process using a small set of ground truth labels. In this work, we introduce AutoWS-Bench-101: a framework for evaluating automated WS (AutoWS) techniques in challenging WS settings---a set of diverse application domains on which it has been previously difficult or impossible to apply traditional WS techniques. While AutoWS is a promising direction toward expanding the application-scope of WS, the emergence of powerful methods such as zero-shot foundation models reveal the need to understand how AutoWS techniques compare or cooperate with modern zero-shot or few-shot learners. This informs the central question of AutoWS-Bench-101: given an initial set of 100 labels for each task, we ask whether a practitioner should use an AutoWS method to generate additional labels or use some simpler baseline, such as zero-shot predictions from a foundation model or supervised learning. We observe that it is necessary for AutoWS methods to incorporate signal from foundation models if they are to outperform simple few-shot baselines, and AutoWS-Bench-101 promotes future research in this direction. We conclude with a thorough ablation study of AutoWS methods.

## [K-Radar: 4D Radar Object Detection for Autonomous Driving in Various Weather Conditions](#)

- Dong-Hee Paek · SEUNG-HYUN KONG · Kevin Tirta Wijaya
- abstract@[open-review](#): Unlike RGB cameras that use visible light bands (384~769 THz) and Lidar that use infrared bands (361~331 THz), Radars use relatively longer wavelength radio bands (77~81 GHz), resulting in robust measurements in adverse weathers. Unfortunately, existing Radar datasets only contain a relatively small number of samples compared to the existing camera and Lidar datasets. This may hinder the development of sophisticated data-driven deep learning techniques for Radar-based perception. Moreover, most of the existing Radar datasets only provide 3D Radar tensor (3DRT) data that contain power measurements along the Doppler, range, and azimuth dimensions. As there is no elevation information, it is challenging to estimate the 3D bounding box of an object from 3DRT. In this work, we introduce KAIST-Radar (K-Radar), a novel large-scale object detection dataset and benchmark that contains 35K frames of 4D Radar tensor (4DRT) data with power measurements along the Doppler, range, azimuth, and elevation dimensions, together with carefully annotated 3D bounding box labels of objects on the roads. K-Radar includes challenging driving conditions such as adverse weathers (fog, rain, and snow) on various road structures (urban, suburban roads, alleyways, and highways). In addition to the 4DRT, we provide auxiliary measurements from carefully calibrated high-resolution Lidars, surround stereo cameras, and RTK-GPS. We also provide 4DRT-based object detection baseline neural networks (baseline NNs) and show that the height information is crucial for 3D object detection. And by comparing the baseline NN with a similarly-structured Lidar-based neural network, we demonstrate that 4D Radar is a more robust sensor for adverse weather conditions. All codes are available at <https://github.com/kaist-avelab/k-radar>.

## [HandMeThat: Human-Robot Communication in Physical and Social Environments](#)

- Yanming Wan · Jiayuan Mao · Josh Tenenbaum
- abstract@[open-review](#): We introduce HandMeThat, a benchmark for a holistic evaluation of instruction understanding and following in physical and social environments. While previous datasets primarily focused on language grounding and planning, HandMeThat considers the resolution of human instructions with ambiguities based on the physical (object states and relations) and social (human actions and goals) information. HandMeThat contains 10,000 episodes of human-robot interactions. In each episode, the robot first observes a trajectory of human actions towards her internal goal. Next, the robot receives a human instruction and should take actions to accomplish the subgoal set through the instruction. In this paper, we present a textual interface for our benchmark, where the robot interacts with a virtual environment through textual commands. We evaluate several baseline models on HandMeThat, and show that both offline and online reinforcement learning algorithms perform poorly on HandMeThat, suggesting significant room for future work on physical and social human-robot communications and interactions.

## [How Well Do Unsupervised Learning Algorithms Model Human Real-time and Life-long Learning?](#)

- Chengxu Zhuang · Ziyu Xiang · Yoon Bai · Xiaoxuan Jia · Nicholas Turk-Browne · Kenneth Norman · James J DiCarlo · Dan Yamins
- abstract@[open-review](#): Humans learn from visual inputs at multiple timescales, both rapidly and flexibly acquiring visual knowledge over short periods, and robustly accumulating online learning progress over longer periods. Modeling these powerful learning capabilities is an important problem for computational visual cognitive science, and models that could replicate them would be of substantial utility in real-world computer vision settings. In this work, we establish benchmarks for both real-time and life-long continual visual learning. Our real-time learning benchmark measures a model's ability to match the rapid visual behavior changes of real humans over the course of minutes and hours, given a stream of visual inputs. Our life-long learning benchmark evaluates the performance of models in a purely online learning curriculum obtained directly from child visual experience over the course of years of development. We evaluate a spectrum of recent deep self-supervised visual learning algorithms on both benchmarks, finding that none of them perfectly match human performance, though some algorithms perform substantially better than others. Interestingly, algorithms embodying recent trends in self-supervised learning -- including BYOL, SwAV and MAE -- are substantially worse on our benchmarks than an earlier generation of self-supervised algorithms such as SimCLR and MoCo-v2. We present analysis indicating that the failure of these newer algorithms is primarily due to their inability to handle the kind of sparse low-diversity datastreams that naturally arise in the real world, and that actively leveraging memory through negative sampling -- a mechanism eschewed by these newer algorithms -- appears useful for facilitating learning in such low-diversity environments. We also illustrate a complementarity between the short and long timescales in the two benchmarks, showing how requiring a single learning algorithm to be locally context-sensitive enough to match real-time learning changes while stable enough to avoid catastrophic forgetting over the long term induces a trade-off that human-like algorithms may have to straddle. Taken together, our benchmarks establish a quantitative way to directly compare learning between neural networks models and human learners, show how choices in the mechanism by which such algorithms handle sample comparison and memory strongly impact their ability to match human learning abilities, and expose an open problem space for identifying more flexible and robust visual self-supervision algorithms.

## [MSDS: A Large-Scale Chinese Signature and Token Digit String Dataset for Handwriting Verification](#)

- Peirong Zhang · Jiajia Jiang · Yuliang Liu · Lianwen Jin
- abstract@[open-review](#): Although online handwriting verification has made great progress recently, the verification performances are still far behind the real usage owing to the small scale of the datasets as well as the limited biometric mediums. Therefore, this paper proposes a new handwriting verification benchmark dataset named Multimodal Signature and Digit String (MSDS), which consists of two subsets: MSDS-ChS (Chinese Signatures) and MSDS-TDS (Token Digit Strings), contributed by 402 users, with 20 genuine samples and 20 skilled forgeries per user per subset. MSDS-ChS consists of handwritten Chinese signatures, which, to the best of our knowledge, is the largest publicly available Chinese signature dataset for handwriting verification, at least eight times larger than existing online datasets. Meanwhile, MSDS-TDS consists of handwritten Token Digit Strings, i.e, the actual phone numbers of users, which have not been explored yet. Extensive experiments with different baselines are respectively conducted for MSDS-ChS and MSDS-TDS. Surprisingly, verification performances of state-of-the-art methods on MSDS-TDS are generally better than those on MSDS-ChS, which indicates that the handwritten Token Digit String could be a more effective biometric than handwritten Chinese signature. This is a promising discovery that could inspire us to explore new biometric traits. The MSDS dataset is available at <https://github.com/HCIILAB/MSDS>.

## [A Unified Evaluation of Textual Backdoor Learning: Frameworks and Benchmarks](#)

- Ganqu Cui · Lifan Yuan · Bingxiang He · Yangyi Chen · Zhiyuan Liu · Maosong Sun
- abstract@[open-review](#): Textual backdoor attacks are a kind of practical threat to NLP systems. By injecting a backdoor in the training phase, the adversary could control model predictions via predefined triggers. As various attack and defense models have been proposed, it is of great significance to perform rigorous evaluations. However, we highlight two issues in previous backdoor learning evaluations: (1) The differences between real-world scenarios (e.g. releasing poisoned datasets or models) are neglected, and we argue that each scenario has its own constraints and concerns, thus requires specific

evaluation protocols; (2) The evaluation metrics only consider whether the attacks could flip the models' predictions on poisoned samples and retain performances on benign samples, but ignore that poisoned samples should also be stealthy and semantic-preserving. To address these issues, we categorize existing works into three practical scenarios in which attackers release datasets, pre-trained models, and fine-tuned models respectively, then discuss their unique evaluation methodologies. On metrics, to completely evaluate poisoned samples, we use grammar error increase and perplexity difference for stealthiness, along with text similarity for validity. After formalizing the frameworks, we develop an open-source toolkit OpenBackdoor to foster the implementations and evaluations of textual backdoor learning. With this toolkit, we perform extensive experiments to benchmark attack and defense models under the suggested paradigm. To facilitate the underexplored defenses against poisoned datasets, we further propose CUBE, a simple yet strong clustering-based defense baseline. We hope that our frameworks and benchmarks could serve as the cornerstones for future model development and evaluations.

## [MBW: Multi-view Bootstrapping in the Wild](#)

- Mosam Dabhi · Chaoyang Wang · Tim Clifford · László Jeni · Ian Fasel · Simon Lucey
- abstract@[open-review](#): Labeling articulated objects in unconstrained settings has a wide variety of applications including entertainment, neuroscience, psychology, ethology, and many fields of medicine. Large offline labeled datasets do not exist for all but the most common articulated object categories (e.g., humans). Hand labeling these landmarks within a video sequence is a laborious task. Learned landmark detectors can help, but can be error-prone when trained from only a few examples. Multi-camera systems that train fine-grained detectors have shown significant promise in detecting such errors, allowing for self-supervised solutions that only need a small percentage of the video sequence to be hand-labeled. The approach, however, is based on calibrated cameras and rigid geometry, making it expensive, difficult to manage, and impractical in real-world scenarios. In this paper, we address these bottlenecks by combining a non-rigid 3D neural prior with deep flow to obtain high-fidelity landmark estimates from videos with only two or three uncalibrated, handheld cameras. With just a few annotations (representing \$1-2\%\$ of the frames), we are able to produce 2D results comparable to state-of-the-art fully supervised methods, along with 3D reconstructions that are impossible with other existing approaches. Our Multi-view Bootstrapping in the Wild (MBW) approach demonstrates impressive results on standard human datasets, as well as tigers, cheetahs, fish, colobus monkeys, chimpanzees, and flamingos from videos captured casually in a zoo. We release the codebase for MBW as well as this challenging zoo dataset consisting of image frames of tail-end distribution categories with their corresponding 2D and 3D labels generated from minimal human intervention.

## [Chartalist: Labeled Graph Datasets for UTXO and Account-based Blockchains](#)

- Kiarash Shamsi · Friedhelm Victor · Murat Kantarcioglu · Yulia Gel · Cuneyt G Akcora
- abstract@[open-review](#): Machine learning on blockchain graphs is an emerging field with many applications such as ransomware payment tracking, price manipulation analysis, and money laundering detection. However, analyzing blockchain data requires domain expertise and computational resources, which pose a significant barrier and hinder advancement in this field. We introduce Chartalist, the first comprehensive platform to methodically access and use machine learning across a large selection of blockchains to address this challenge. Chartalist contains ML-ready datasets from unspent transaction output (UTXO) (e.g., Bitcoin) and account-based blockchains (e.g., Ethereum). We envision that Chartalist can facilitate data modeling, analysis, and representation of blockchain data and attract a wider community of scientists to analyze blockchains. Chartalist is an open-science initiative at <https://github.com/cakcora/Chartalist>.

## [Learning Long-Term Crop Management Strategies with CyclesGym](#)

- Matteo Turchetta · Luca Corinzia · Scott Sussex · Amanda Burton · Juan Herrera · Ioannis Athanasiadis · Joachim M Buhmann · Andreas Krause
- abstract@[open-review](#): To improve the sustainability and resilience of modern food systems, designing improved crop management strategies is crucial. The increasing abundance of data on agricultural systems suggests that future strategies could benefit from adapting to environmental conditions, but how to design these adaptive policies poses a new frontier. A natural technique for learning policies in these kinds of sequential decision-making problems is reinforcement learning (RL). To obtain the large number of samples required to learn effective RL policies, existing work has used mechanistic crop growth models (CGMs) as simulators. These solutions focus on single-year, single-crop simulations for learning strategies for a single agricultural management practice. However, to learn sustainable long-term policies we must be able to train in multi-year environments, with multiple crops, and consider a wider array of management techniques. We introduce CYCLESGYM, an RL environment based on the multi-year, multi-crop CGM Cycles. CYCLESGYM allows for long-term planning in agroecosystems, provides modular state space and reward constructors and weather generators, and allows for complex actions. For RL researchers, this is a novel benchmark to investigate issues arising in real-world applications. For agronomists, we demonstrate the potential of RL as a powerful optimization tool for agricultural systems management in multi-year case studies on nitrogen (N) fertilization and crop planning scenarios.

## [PDEBench: An Extensive Benchmark for Scientific Machine Learning](#)

- Makoto Takamoto · Timothy Praditia · Raphael Leiteritz · Daniel MacKinlay · Francesco Alesiani · Dirk Pflüger · Mathias Niepert
- abstract@[open-review](#): Machine learning-based modeling of physical systems has experienced increased interest in recent years. Despite some impressive progress, there is still a lack of benchmarks for Scientific ML that are easy to use but still challenging and representative of a wide range of problems. We introduce PDEBENCH, a benchmark suite of time-dependent simulation tasks based on Partial Differential Equations (PDEs). PDEBENCH comprises both code and data to benchmark the performance of novel machine learning models against both classical numerical simulations and machine learning baselines. Our proposed set of benchmark problems contribute the following unique features: (1) A much wider range of PDEs compared to existing benchmarks, ranging from relatively common examples to more realistic and difficult problems; (2) much larger ready-to-use datasets compared to prior work, comprising multiple simulation runs across a larger number of initial and boundary conditions and PDE parameters; (3) more extensible source codes with user-friendly APIs for data generation and baseline results with popular machine learning models (FNO, U-Net, PINN, Gradient-Based Inverse Method). PDEBENCH allows researchers to extend the benchmark freely for their own purposes using a standardized API and to compare the performance of new models to existing baseline methods. We also propose new evaluation metrics with the aim to provide a more holistic understanding of learning methods in the context of Scientific ML. With those metrics we identify tasks which are challenging for recent ML methods and propose these tasks as future challenges for the community. The code is available at <https://github.com/pdebench/PDEBench>.

## [Unravelling the Performance of Physics-informed Graph Neural Networks for Dynamical Systems](#)

- Abishek Thangamuthu · Gunjan Kumar · Suresh Bishnoi · Ravinder Bhattoo · N M Anoop Krishnan · Sayan Ranu
- abstract@[open-review](#): Recently, graph neural networks have been gaining a lot of attention to simulate dynamical systems due to their inductive nature leading to zero-shot generalizability. Similarly, physics-informed inductive biases in deep-learning frameworks have been shown to give superior performance in learning the dynamics of physical systems. There is a growing volume of literature that attempts to combine these two approaches. Here, we evaluate the performance of thirteen different graph neural networks, namely, Hamiltonian and Lagrangian graph neural networks, graph neural ODE, and their variants with explicit constraints and different architectures. We briefly explain the theoretical formulation highlighting the similarities and differences in the inductive biases and graph architecture of these systems. Then, we evaluate them on spring, pendulum, and gravitational and 3D deformable solid systems to compare the performance in terms of rollout error, conserved quantities such as energy and momentum, and generalizability to unseen system sizes. Our study demonstrates that GNNs with additional inductive biases, such as explicit constraints and decoupling of kinetic and potential energies, exhibit significantly enhanced performance. Further, all the physics-informed GNNs exhibit zero-shot generalizability to system sizes an order of magnitude larger than the training system, thus providing a promising route to simulate large-scale realistic systems.

## [A Greek Parliament Proceedings Dataset for Computational Linguistics and Political Analysis](#)

- Konstantina Dritsa · Aikaterini Thoma · Ioannis Pavlopoulos · Panos Louridas
- abstract@[open-review](#): Large, diachronic datasets of political discourse are hard to come across, especially for resource-lean languages such as Greek. In this paper, we introduce a curated dataset of the Greek Parliament Proceedings that extends chronologically from 1989 up to 2020. It consists of more than 1 million speeches with extensive meta-data, extracted from 5,355 parliamentary sitting record files. We explain how it was constructed and the challenges that had to be overcome. The dataset can be used for both computational linguistics and political analysis---ideally, combining the two. We present such an application, showing (i) how the dataset can be used to study the change of word usage through time, (ii) between significant historical events and political parties, (iii) by evaluating and employing algorithms for detecting semantic shifts.

## [A Survey and Datasheet Repository of Publicly Available US Criminal Justice Datasets](#)

- Miri Zilka · Bradley Butcher · Adrian Weller
- abstract@[open-review](#): Criminal justice is an increasingly important application domain for machine learning and algorithmic fairness, as predictive tools are becoming widely used in police, courts, and prison systems worldwide. A few relevant benchmarks have received significant attention, e.g., the COMPAS dataset, often without proper consideration of the domain context. To raise awareness of publicly available criminal justice datasets and encourage their responsible use, we conduct a survey, consider contexts, highlight potential uses, and identify gaps and limitations. We provide datasheets for 15 datasets and upload them to a public repository. We compare the datasets across several dimensions, including size, coverage of the population, and potential use, highlighting concerns. We hope that this work can provide a useful starting point for researchers looking for appropriate datasets related to criminal justice, and that the repository will continue to grow as a community effort.

## [CARLANE: A Lane Detection Benchmark for Unsupervised Domain Adaptation from Simulation to multiple Real-World Domains](#)

- Bonifaz Stuhr · Johann Haselberger · Julian Gebele
- abstract@[open-review](#): Unsupervised Domain Adaptation demonstrates great potential to mitigate domain shifts by transferring models from labeled source domains to unlabeled target domains. While Unsupervised Domain Adaptation has been applied to a wide variety of complex vision tasks, only few works focus on lane detection for autonomous driving. This can be attributed to the lack of publicly available datasets. To facilitate research in these directions, we propose CARLANE, a 3-way sim-to-real domain adaptation benchmark for 2D lane detection. CARLANE encompasses the single-target datasets MoLane and TuLane and the multi-target dataset MuLane. These datasets are built from three different domains, which cover diverse scenes and contain a total of 163K unique images, 118K of which are annotated. In addition we evaluate and report systematic baselines, including our own method, which builds upon Prototypical Cross-domain Self-supervised Learning. We find that false positive and false negative rates of the evaluated domain adaptation methods are high compared to those of fully supervised baselines. This affirms the need for benchmarks such as CARLANE to further strengthen research in Unsupervised Domain Adaptation for lane detection. CARLANE, all evaluated models and the corresponding implementations are publicly available at <https://carlanebenchmark.github.io>.

## [SurDis: A Surface Discontinuity Dataset for Wearable Technology to Assist Blind Navigation in Urban Environments](#)

- Kuan Yew Leong · Siew Mooi Lim
- abstract@[open-review](#): According to World Health Organization, there is an estimated 2.2 billion people with a near or distance vision impairment worldwide. Difficulty in self-navigation is one of the greatest challenges to independence for the blind and low vision (BLV) people. Through consultations with several BLV service providers, we realized that negotiating surface discontinuities is one of the very prominent challenges when navigating an outdoor environment within the urban. Surface discontinuities are commonly formed by rises and drop-offs along a pathway. They could be a threat to balancing during a walk and perceiving such a threat is highly challenging to the BLVs. In this paper, we introduce SurDis, a novel dataset of depth maps and stereo images that exemplifies the issue of surface discontinuity in the urban areas of Klang Valley, Malaysia. We seek to address the limitation of existing datasets of such nature in these areas. Current mobility tools for the BLVs predominantly focus on furniture, indoor built environments, traffic signs, vehicles, humans and various types of objects' detection above the surface of a pathway. We emphasize a specific purpose for SurDis --- to support the development of assistive wearable technology for the BLVs to negotiate surface discontinuity. We consulted BLV volunteers on the specifications of surface condition that could become hazardous for navigation using 3D printed replicas of actual scaled-down scenes, and identified locations that are frequented by the BLVs as our target data collection fields. With feedback from these volunteers, we developed a lightweight, small and unobtrusive prototype equipped with a tiny stereo camera and an embedded system on a single board computer to capture the samples from 10 different locations. We describe instrument development, data collection, preprocessing, annotation, and experiments conducted. The dataset contains: (1) more than 17000 depth maps generated from 200 sets of stereo image sequences, (2) annotations of surface discontinuity in the depth maps, and (3) bitmap stereo image pairs corresponding to the depth maps in (1).

## [Towards Open Set 3D Learning: Benchmarking and Understanding Semantic Novelty Detection on Pointclouds](#)

- Antonio Alliegro · Francesco Cappio Borlino · Tatiana Tommasi
- abstract@[open-review](#): In recent years there has been significant progress in the field of 3D learning on classification, detection and segmentation problems. The vast majority of the existing studies focus on canonical closed-set conditions, neglecting the intrinsic open nature of the real-world. This limits the abilities of robots and autonomous systems involved in safety-critical applications that require managing novel and unknown signals. In this context exploiting 3D data can be a valuable asset since it provides rich information about the geometry of sensed objects and scenes. With this paper we provide the first broad study on Open Set 3D learning. We introduce a novel testbed for semantic novelty detection that considers several settings with increasing difficulties in terms of category semantic shift, and covers both in-domain (synthetic-to-synthetic, real-to-real) and cross-domain (synthetic- to-real) scenarios. Moreover, we investigate the related Open Set 2D literature to understand if and how its recent improvements are effective on 3D data. Our extensive benchmark positions several algorithms in the same coherent picture, revealing their strengths and limitations. The results of our analysis may serve as a reliable foothold for future tailored Open Set 3D models.

## [OLIVES Dataset: Ophthalmic Labels for Investigating Visual Eye Semantics](#)

- Mohit Prabhushankar · Kiran Kokilepersaud · Yash-yeo Logan · Stephanie Trejo Corona · Ghassan AlRegib · Charles Wykoff
- abstract@[open-review](#): Clinical diagnosis of the eye is performed over multifarious data modalities including scalar clinical labels, vectorized biomarkers, two-dimensional fundus images, and three-dimensional Optical Coherence Tomography (OCT) scans. Clinical practitioners use all available data modalities for diagnosing and treating eye diseases like Diabetic Retinopathy (DR) or Diabetic Macular Edema (DME). Enabling usage of machine learning algorithms within the ophthalmic medical domain requires research into the relationships and interactions between all relevant data over a treatment period. Existing datasets are limited in that they neither provide data nor consider the explicit relationship modeling between the data modalities. In this paper, we introduce the Ophthalmic Labels for Investigating Visual Eye Semantics (OLIVES) dataset that addresses the above limitation. This is the first OCT and near-IR fundus dataset that includes clinical labels, biomarker labels, disease labels, and time-series patient treatment information from associated clinical trials. The dataset consists of 1268 near-IR fundus images each with at least 49 OCT scans, and 16 biomarkers, along with 4 clinical labels and a disease diagnosis of DR or DME. In total, there are 96 eyes' data averaged over a period of at least two years with each eye treated for an average of 66 weeks and 7 injections. We benchmark the utility of OLIVES dataset for ophthalmic data as well as provide benchmarks and concrete research directions for core and emerging machine learning paradigms within medical image analysis.

## [A Benchmark for Compositional Visual Reasoning](#)

- Aimen Zerroug · Mohit Vaishnav · Julien Colin · Sebastian Musslick · Thomas Serre
- abstract@[open-review](#): A fundamental component of human vision is our ability to parse complex visual scenes and judge the relations between their constituent objects. AI benchmarks for visual reasoning have driven rapid progress in recent years with state-of-the-art systems now reaching human accuracy on some of these benchmarks. Yet, there remains a major gap between humans and AI systems in terms of the sample efficiency with which they learn new visual reasoning tasks. Humans' remarkable efficiency at learning has been at least partially attributed to their ability to harness compositionality -- allowing them to efficiently take advantage of previously gained knowledge when learning new tasks. Here, we introduce a novel visual reasoning benchmark, Compositional Visual Relations (CVR), to drive progress towards the development of more data-efficient learning algorithms. We take inspiration from fluidic intelligence and non-verbal reasoning tests and describe a novel method for creating compositions of abstract rules and generating image datasets corresponding to these rules at scale. Our proposed benchmark includes measures of sample efficiency, generalization, compositionality, and transfer across task rules. We systematically evaluate modern neural architectures and find that convolutional architectures surpass transformer-based architectures across all performance measures in most data regimes. However, all computational models are much less data efficient than humans, even after learning informative visual representations using self-supervision. Overall, we hope our challenge will spur interest in developing neural architectures that can learn to harness compositionality for more efficient learning.

## [pyKT: A Python Library to Benchmark Deep Learning based Knowledge Tracing Models](#)

- Zitao Liu · Qiongqiong Liu · Jiahao Chen · Shuyan Huang · Jiliang Tang · Weiqi Luo
- abstract@[open-review](#): Knowledge tracing (KT) is the task of using students' historical learning interaction data to model their knowledge mastery over time so as to make predictions on their future interaction performance. Recently, remarkable progress has been made of using various deep learning techniques to solve the KT problem. However, the success behind deep learning based knowledge tracing (DLKT) approaches is still left somewhat unknown and proper measurement and analysis of these DLKT approaches remain a challenge. First, data preprocessing procedures in existing works are often private and custom, which limits experimental standardization. Furthermore, existing DLKT studies often differ in terms of the evaluation protocol and are far away real-world educational contexts. To address these problems, we introduce a comprehensive python based benchmark platform, \textsc{pyKT}, to guarantee valid comparisons across DLKT methods via thorough evaluations. The \textsc{pyKT} library consists of a standardized set of integrated data preprocessing procedures on 7 popular datasets across different domains, and 10 frequently compared DLKT model implementations for transparent experiments. Results from our fine-grained and rigorous empirical KT studies yield a set of observations and suggestions for effective DLKT, e.g., wrong evaluation setting may cause label leakage that generally leads to performance inflation; and the improvement of many DLKT approaches is minimal compared to the very first DLKT model proposed by Piech et al. \cite{piech2015deep}. We have open sourced \textsc{pyKT} and our experimental results at \url{https://pykt.org/}. We welcome contributions from other research groups and practitioners.

## [xView3-SAR: Detecting Dark Fishing Activity Using Synthetic Aperture Radar Imagery](#)

- Fernando Paolo · Tsu-ting Tim Lin · Ritwik Gupta · Bryce Goodman · Nirav Patel · Daniel Kuster · David Kroodsma · Jared Dunnmon
- abstract@[open-review](#): Unsustainable fishing practices worldwide pose a major threat to marine resources and ecosystems. Identifying vessels that do not show up in conventional monitoring systems--known as ``dark vessels''--is key to managing and securing the health of marine environments. With the rise of satellite-based synthetic aperture radar (SAR) imaging and modern machine learning (ML), it is now possible to automate detection of dark vessels day or night, under all-weather conditions. SAR images, however, require a domain-specific treatment and are not widely accessible to the ML community. Maritime objects (vessels and offshore infrastructure) are relatively small and sparse, challenging traditional computer vision approaches. We present the largest labeled dataset for training ML models to detect and characterize vessels and ocean structures in SAR imagery. xView3-SAR consists of nearly 1,000 analysis-ready SAR images from the Sentinel-1 mission that are, on average, 29,400-by-24,400 pixels each. The images are annotated using a combination of automated and manual analysis. Co-located bathymetry and wind state rasters accompany every SAR image. We also provide an overview of the xView3 Computer Vision Challenge, an international competition using xView3-SAR for ship detection and characterization at large scale. We release the data (\url{https://iuu.xview.us/}\{https://iuu.xview.us/}) and code (\url{https://github.com/DIUx-xView}\{https://github.com/DIUx-xView}) to support ongoing development and evaluation of ML approaches for this important application.

## [ETAB: A Benchmark Suite for Visual Representation Learning in Echocardiography](#)

- Ahmed M. Alaa · Anthony Philippakis · David Sontag
- abstract@[open-review](#): Echocardiography is one of the most commonly used diagnostic imaging modalities in cardiology. Application of deep learning models to echocardiograms can enable automated identification of cardiac structures, estimation of cardiac function, and prediction of clinical outcomes. However, a major hindrance to realizing the full potential of deep learning is the lack of large-scale, fully curated and annotated data sets required for supervised training. High-quality pre-trained representations that can transfer useful visual features of echocardiograms to downstream tasks can help adapt deep learning models to new setups using fewer examples. In this paper, we design a suite of benchmarks that can be used to pre-train and evaluate echocardiographic representations with respect to various clinically-relevant tasks using publicly accessible data sets. In addition, we develop a unified evaluation protocol---which we call the echocardiographic task adaptation benchmark (ETAB)---that measures how well a visual representation of echocardiograms generalizes to common downstream tasks of interest. We use our benchmarking framework to evaluate state-of-the-art vision modeling pipelines. We envision that our standardized, publicly accessible benchmarks would encourage future research and expedite progress in applying deep learning to high-impact problems in cardiovascular medicine.

## [The BigScience ROOTS Corpus: A 1.6TB Composite Multilingual Dataset](#)

- Hugo Laurençon · Lucile Saulnier · Thomas Wang · Christopher Akiki · Albert Villanova del Moral · Teven Le Scao · Leandro Von Werra · Chenghao Mou · Eduardo González Ponferrada · Huu Nguyen · Jørg Frohberg · Mario Álvarez · Quentin Lhoest · Angelina McMillan-Major · Gerard Dupont · Stella Biderman · Anna Rogers · Loubna Ben allal · Francesco De Toni · Giada Pistilli · Olivier Nguyen · Somaieh Nikpoor · Maraim Masoud · Pierre Colombo · Javier de la Rosa · Paulo Villegas · Tristan Thrush · Shayne Longpre · Sebastian Nagel · Leon Weber · Manuel Muñoz · Jian Zhu · Daniel Van Strien · Zaid Alyafeai · Khalid Almubarak · Minh Chien Vu · Itziar Gonzalez-Dios · Aitor Soroa · Kyle Lo · Manan Dey · Pedro Ortiz Suarez · Aaron Gokaslan · Shamik Bose · David Adelani · Long Phan · Hieu Tran · Ian Yu · Suhas Pai · Jenny Chim · Violette Lepercq · Suzana Ilic · Margaret Mitchell · Alexandra V Luccioni · Yacine Jernite
- abstract@[open-review](#): As language models grow ever larger, the need for large-scale high-quality text datasets has never been more pressing, especially in multilingual settings. The BigScience workshop, a 1-year international and multidisciplinary initiative, was formed with the goal of researching and training large language models as a values-driven undertaking, putting issues of ethics, harm, and governance in the foreground. This paper documents the data creation and curation efforts undertaken by BigScience to assemble the Responsible Open-science Open-collaboration Text Sources (ROOTS) corpus, a 1.6TB dataset spanning 59 languages that was used to train the 176-billion-parameter BigScience Large Open-science Open-access Multilingual (BLOOM) language model. We further release a large initial subset of the corpus and analyses thereof, and hope to empower large-scale monolingual and multilingual modeling projects with both the data and the processing tools, as well as stimulate research around this large multilingual corpus.

## [APT-36K: A Large-scale Benchmark for Animal Pose Estimation and Tracking](#)

- Yuxiang Yang · Junjie Yang · Yufei Xu · Jing Zhang · Long Lan · Dacheng Tao

- abstract@[open-review](#): Animal pose estimation and tracking (APT) is a fundamental task for detecting and tracking animal keypoints from a sequence of video frames. Previous animal-related datasets focus either on animal tracking or single-frame animal pose estimation, and never on both aspects. The lack of APT datasets hinders the development and evaluation of video-based animal pose estimation and tracking methods, limiting the applications in real world, e.g., understanding animal behavior in wildlife conservation. To fill this gap, we make the first step and propose APT-36K, i.e., the first large-scale benchmark for animal pose estimation and tracking. Specifically, APT-36K consists of 2,400 video clips collected and filtered from 30 animal species with 15 frames for each video, resulting in 36,000 frames in total. After manual annotation and careful double-check, high-quality keypoint and tracking annotations are provided for all the animal instances. Based on APT-36K, we benchmark several representative models on the following three tracks: (1) supervised animal pose estimation on a single frame under intra- and inter-domain transfer learning settings, (2) inter-species domain generalization test for unseen animals, and (3) animal pose estimation with animal tracking. Based on the experimental results, we gain some empirical insights and show that APT-36K provides a useful animal pose estimation and tracking benchmark, offering new challenges and opportunities for future research. The code and dataset will be made publicly available at <https://github.com/pandorgan/APT-36K>.

## [OpenFWI: Large-scale Multi-structural Benchmark Datasets for Full Waveform Inversion](#)

- Chengyuan Deng · Shihang Feng · Hanchen Wang · Xitong Zhang · Peng Jin · Yinan Feng · Qili Zeng · Yinpeng Chen · Youzuo Lin
- abstract@[open-review](#): Full waveform inversion (FWI) is widely used in geophysics to reconstruct high-resolution velocity maps from seismic data. The recent success of data-driven FWI methods results in a rapidly increasing demand for open datasets to serve the geophysics community. We present OpenFWI, a collection of large-scale multi-structural benchmark datasets, to facilitate diversified, rigorous, and reproducible research on FWI. In particular, OpenFWI consists of \$12\$ datasets (\$2.1\$TB in total) synthesized from multiple sources. It encompasses diverse domains in geophysics (interface, fault, CO\$\_2\$ reservoir, etc.), covers different geological subsurface structures (flat, curve, etc.), and contain various amounts of data samples (2K - 67K). It also includes a dataset for 3D FWI. Moreover, we use OpenFWI to perform benchmarking over four deep learning methods, covering both supervised and unsupervised learning regimes. Along with the benchmarks, we implement additional experiments, including physics-driven methods, complexity analysis, generalization study, uncertainty quantification, and so on, to sharpen our understanding of datasets and methods. The studies either provide valuable insights into the datasets and the performance, or uncover their current limitations. We hope OpenFWI supports prospective research on FWI and inspires future open-source efforts on AI for science. All datasets and related information can be accessed through our website at <https://openfwi-lanl.github.io/>

## [Forecasting Future World Events With Neural Networks](#)

- Andy Zou · Tristan Xiao · Ryan Jia · Joe Kwon · Mantas Mazeika · Richard Li · Dawn Song · Jacob Steinhardt · Owain Evans · Dan Hendrycks
- abstract@[open-review](#): Forecasting future world events is a challenging but valuable task. Forecasts of climate, geopolitical conflict, pandemics and economic indicators help shape policy and decision making. In these domains, the judgment of expert humans contributes to the best forecasts. Given advances in language modeling, can these forecasts be automated? To this end, we introduce Autocast, a dataset containing thousands of forecasting questions and an accompanying news corpus. Questions are taken from forecasting tournaments, ensuring high quality, real-world importance, and diversity. The news corpus is organized by date, allowing us to precisely simulate the conditions under which humans made past forecasts (avoiding leakage from the future). Motivated by the difficulty of forecasting numbers across orders of magnitude (e.g. global cases of COVID-19 in 2022), we also curate IntervalQA, a dataset of numerical questions and metrics for calibration. We test language models on our forecasting task and find that performance is far below a human expert baseline. However, performance improves with increased model size and incorporation of relevant information from the news corpus. In sum, Autocast poses a novel challenge for large language models and improved performance could bring large practical benefits.

## [How Would The Viewer Feel? Estimating Wellbeing From Video Scenarios](#)

- Mantas Mazeika · Eric Tang · Andy Zou · Steven Basart · Jun Shern Chan · Dawn Song · David Forsyth · Jacob Steinhardt · Dan Hendrycks
- abstract@[open-review](#): In recent years, deep neural networks have demonstrated increasingly strong abilities to recognize objects and activities in videos. However, as video understanding becomes widely used in real-world applications, a key consideration is developing human-centric systems that understand not only the content of the video but also how it would affect the wellbeing and emotional state of viewers. To facilitate research in this setting, we introduce two large-scale datasets with over 60,000 videos manually annotated for emotional response and subjective wellbeing. The Video Cognitive Empathy (VCE) dataset contains annotations for distributions of fine-grained emotional responses, allowing models to gain a detailed understanding of affective states. The Video to Valence (V2V) dataset contains annotations of relative pleasantness between videos, which enables predicting a continuous spectrum of wellbeing. In experiments, we show how video models that are primarily trained to recognize actions and find contours of objects can be repurposed to understand human preferences and the emotional content of videos. Although there is room for improvement, predicting wellbeing and emotional response is on the horizon for state-of-the-art models. We hope our datasets can help foster further advances at the intersection of commonsense video understanding and human preference learning.

## [StrokeRehab: A Benchmark Dataset for Sub-second Action Identification](#)

- Aakash Kaku · Kangning Liu · Avinash Parnandi · Haresh Rengaraj Rajamohan · Kannan Venkataramanan · Anita Venkatesan · Audre Wirtanen · Natasha Pandit · Heidi Schambra · Carlos Fernandez-Granda
- abstract@[open-review](#): Automatic action identification from video and kinematic data is an important machine learning problem with applications ranging from robotics to smart health. Most existing works focus on identifying coarse actions such as running, climbing, or cutting vegetables, which have relatively long durations and a complex series of motions. This is an important limitation for applications that require identification of more elemental motions at high temporal resolution. For example, in the rehabilitation of arm impairment after stroke, quantifying the training dose (number of repetitions) requires differentiating motions with sub-second durations. Our goal is to bridge this gap. To this end, we introduce a large-scale, multimodal dataset, StrokeRehab, as a new action-recognition benchmark that includes elemental short-duration actions labeled at a high temporal resolution. StrokeRehab consists of a high-quality inertial measurement unit sensor and video data of 51 stroke-impaired patients and 20 healthy subjects performing activities of daily living like feeding, brushing teeth, etc. Because it contains data from both healthy and impaired individuals, StrokeRehab can be used to study the influence of distribution shift in action-recognition tasks. When evaluated on StrokeRehab, current state-of-the-art models for action segmentation produce noisy predictions, which reduces their accuracy in identifying the corresponding sequence of actions. To address this, we propose a novel approach for high-resolution action identification, inspired by speech-recognition techniques, which is based on a sequence-to-sequence model that directly predicts the sequence of actions. This approach outperforms current state-of-the-art methods on StrokeRehab, as well as on the standard benchmark datasets 50Salads, Breakfast, and Jigsaws.

## [BOND: Benchmarking Unsupervised Outlier Node Detection on Static Attributed Graphs](#)

- Kay Liu · Yingtong Dou · Yue Zhao · Xueying Ding · Xiyang Hu · Ruitong Zhang · Kaize Ding · Canyu Chen · Hao Peng · Kai Shu · Lichao Sun · Jundong Li · George H Chen · Zhihao Jia · Philip S Yu
- abstract@[open-review](#): Detecting which nodes in graphs are outliers is a relatively new machine learning task with numerous applications. Despite the proliferation of algorithms developed in recent years for this task, there has been no standard comprehensive setting for performance evaluation. Consequently, it has been difficult to understand which methods work well and when under a broad range of settings. To bridge this gap, we present "to the best of our knowledge" the first comprehensive benchmark for unsupervised outlier node detection on static attributed graphs called BOND, with the following highlights. (1) We benchmark the outlier detection performance of 14 methods ranging from classical matrix factorization to the latest graph

neural networks. (2) Using nine real datasets, our benchmark assesses how the different detection methods respond to two major types of synthetic outliers and separately to “organic” (real non-synthetic) outliers. (3) Using an existing random graph generation technique, we produce a family of synthetic datasets of different graph sizes that enable us to compare the running time and memory usage of the different outlier detection algorithms as a function of graph size. Based on our experimental results, we discuss the pros and cons of existing graph outlier detection algorithms, and we highlight opportunities for future research. Importantly, our code is freely available and meant to be easily extendable: <https://github.com/pygod-team/pygod/tree/main/benchmark>

## [TwiBot-22: Towards Graph-Based Twitter Bot Detection](#)

- Shangbin Feng · Zhaoxuan Tan · Herun Wan · Ningnan Wang · Zilong Chen · Binchi Zhang · Qinghua Zheng · Wenqian Zhang · Zhenyu Lei · Shujie Yang · Xinshun Feng · Qingyue Zhang · Hongrui Wang · Yuhan Liu · Yuyang Bai · Heng Wang · Zijian Cai · Yanbo Wang · Lijing Zheng · Zihan Ma · Jundong Li · Minnan Luo
- abstract@[open-review](#): Twitter bot detection has become an increasingly important task to combat misinformation, facilitate social media moderation, and preserve the integrity of the online discourse. State-of-the-art bot detection methods generally leverage the graph structure of the Twitter network, and they exhibit promising performance when confronting novel Twitter bots that traditional methods fail to detect. However, very few of the existing Twitter bot detection datasets are graph-based, and even these few graph-based datasets suffer from limited dataset scale, incomplete graph structure, as well as low annotation quality. In fact, the lack of a large-scale graph-based Twitter bot detection benchmark that addresses these issues has seriously hindered the development and evaluation of novel graph-based bot detection approaches. In this paper, we propose TwiBot-22, a comprehensive graph-based Twitter bot detection benchmark that presents the largest dataset to date, provides diversified entities and relations on the Twitter network, and has considerably better annotation quality than existing datasets. In addition, we re-implement 35 representative Twitter bot detection baselines and evaluate them on 9 datasets, including TwiBot-22, to promote a fair comparison of model performance and a holistic understanding of research progress. To facilitate further research, we consolidate all implemented codes and datasets into the TwiBot-22 evaluation framework, where researchers could consistently evaluate new models and datasets. The TwiBot-22 Twitter bot detection benchmark and evaluation framework are publicly available at \url{https://twibot22.github.io/}.

## [FLamby: Datasets and Benchmarks for Cross-Silo Federated Learning in Realistic Healthcare Settings](#)

- Jean Ogier du Terrail · Samy-Safwan Ayed · Edwige Cyffers · Felix Grimberg · Chaoyang He · Regis Loeb · Paul Mangold · Tanguy Marchand · Othmane Marfoq · Erum Mushtaq · Boris Muzellec · Constantin Philippenko · Santiago Silva · Maria Teleczuk · Shadi Albarqouni · Salman Avestimehr · Aurélien Bellet · Aymeric Dieuleveut · Martin Jaggi · Sai Praneeth Karimireddy · Marco Lorenzi · Giovanni Neglia · Marc Tommasi · Mathieu Andreux
- abstract@[open-review](#): Federated Learning (FL) is a novel approach enabling several clients holding sensitive data to collaboratively train machine learning models, without centralizing data. The cross-silo FL setting corresponds to the case of few (\$2--\$50\$) reliable clients, each holding medium to large datasets, and is typically found in applications such as healthcare, finance, or industry. While previous works have proposed representative datasets for cross-device FL, few realistic healthcare cross-silo FL datasets exist, thereby slowing algorithmic research in this critical application. In this work, we propose a novel cross-silo dataset suite focused on healthcare, FLaMby (Federated Learning AMple Benchmark of Your cross-silo strategies), to bridge the gap between theory and practice of cross-silo FL. FLaMby encompasses 7 healthcare datasets with natural splits, covering multiple tasks, modalities, and data volumes, each accompanied with baseline training code. As an illustration, we additionally benchmark standard FL algorithms on all datasets. Our flexible and modular suite allows researchers to easily download datasets, reproduce results and re-use the different components for their research. FLaMby is available at \url{www.github.com/owkin/flamby}.

## [TAP-Vid: A Benchmark for Tracking Any Point in a Video](#)

- Carl Doersch · Ankush Gupta · Larisa Markeeva · Adria Recasens · Lucas Smaira · Yusuf Aytar · Joao Carreira · Andrew Zisserman · Yi Yang
- abstract@[open-review](#): Generic motion understanding from video involves not only tracking objects, but also perceiving how their surfaces deform and move. This information is useful to make inferences about 3D shape, physical properties and object interactions. While the problem of tracking arbitrary physical points on surfaces over longer video clips has received some attention, no dataset or benchmark for evaluation existed, until now. In this paper, we first formalize the problem, naming it tracking any point (TAP). We introduce a companion benchmark, TAP-Vid, which is composed of both real-world videos with accurate human annotations of point tracks, and synthetic videos with perfect ground-truth point tracks. Central to the construction of our benchmark is a novel semi-automatic crowdsourced pipeline which uses optical flow estimates to compensate for easier, short-term motion like camera shake, allowing annotators to focus on harder sections of the video. We validate our pipeline on synthetic data and propose a simple end-to-end point tracking model, TAP-Net, showing that it outperforms all prior methods on our benchmark when trained on synthetic data.

## [EHRSQ: A Practical Text-to-SQL Benchmark for Electronic Health Records](#)

- GYUBOK LEE · Hyeonji Hwang · Seongsu Bae · Yeonsu Kwon · Woncheol Shin · Seongjun Yang · Minjoon Seo · Jong-Yeup Kim · Edward Choi
- abstract@[open-review](#): We present a new text-to-SQL dataset for electronic health records (EHRs), where the utterances are collected from 222 hospital staff—including physicians, nurses, and insurance review and health records teams through a poll conducted at a university hospital. To construct a QA dataset on structured EHR data, we templated the utterances and manually linked them to two open-source EHR databases—MIMIC-III and eICU—along with various time expressions and held-out unanswerable questions, which were all collected from the poll. Our dataset poses a unique set of challenges: the model needs to 1) generate SQL queries that reflect a wide range of needs in the hospital, including simple retrieval and complex operations such as calculating survival rate, 2) understand various time expressions to answer time-sensitive questions in healthcare, and 3) distinguish whether a given question is answerable or unanswerable based on the prediction confidence. We believe our dataset, EHRSQ, could serve as a practical benchmark to develop and assess QA models on structured EHR data and take one step further towards bridging the gap between text-to-SQL research and its real-life deployment in healthcare.

## [FinRL-Meta: Market Environments and Benchmarks for Data-Driven Financial Reinforcement Learning](#)

- Xiao-Yang Liu · Ziyi Xia · Jingyang Rui · Jiechao Gao · Hongyang Yang · Ming Zhu · Christina Wang · Zhaoran Wang · Jian Guo
- abstract@[open-review](#): Finance is a particularly difficult playground for deep reinforcement learning. However, establishing high-quality market environments and benchmarks on financial reinforcement learning are challenging due to three major factors, namely, low signal-to-noise ratio of financial data, survivorship bias of historical data, and information leakage in the backtesting stage. In this paper, we present an openly accessible FinRL-Meta library that has been actively maintained by the FinRL community. First, following a DataOps paradigm, we provide hundreds of market environments through an automatic pipeline that collects dynamic datasets from real-world markets and processes them into standard gym-style market environments. Second, we benchmark popular papers as stepping stones for users to design new trading strategies. We also hold our benchmarks on cloud platforms so that users can visualize their own results and assess the relative performance via community-wise competitions. Third, FinRL-Meta provides tens of Jupyter/Python demos organized in a curriculum and a documentation website to serve the rapidly growing community. FinRL-Meta is available at: \url{https://github.com/AI4Finance-Foundation/FinRL-Meta}.

## [Is one annotation enough? - A data-centric image classification benchmark for noisy and ambiguous label estimation](#)

- Lars Schmarje Å· Vasco Grossmann Å· Claudius Zelenka Å· Sabine Dippel Å· Rainer Kiko Å· Mariusz Oszust Å· Matti Pastell Å· Jenny Stracke Å· Anna Valros Å· Nina Volkmann Å· Reinhard Koch
- abstract@[open-review](#): High-quality data is necessary for modern machine learning. However, the acquisition of such data is difficult due to noisy and ambiguous annotations of humans. The aggregation of such annotations to determine the label of an image leads to a lower data quality. We propose a data-centric image classification benchmark with nine real-world datasets and multiple annotations per image to allow researchers to investigate and quantify the impact of such data quality issues. With the benchmark we can study the impact of annotation costs and (semi-)supervised methods on the data quality for image classification by applying a novel methodology to a range of different algorithms and diverse datasets. Our benchmark uses a two-phase approach via a data label improvement method in the first phase and a fixed evaluation model in the second phase. Thereby, we give a measure for the relation between the input labeling effort and the performance of (semi-)supervised algorithms to enable a deeper insight into how labels should be created for effective model training. Across thousands of experiments, we show that one annotation is not enough and that the inclusion of multiple annotations allows for a better approximation of the real underlying class distribution. We identify that hard labels can not capture the ambiguity of the data and this might lead to the common issue of overconfident models. Based on the presented datasets, benchmarked methods, and analysis, we create multiple research opportunities for the future directed at the improvement of label noise estimation approaches, data annotation schemes, realistic (semi-)supervised learning, or more reliable image collection.

## [Wukong: A 100 Million Large-scale Chinese Cross-modal Pre-training Benchmark](#)

- Jiaxi Gu Å· Xiaojun Meng Å· Guansong Lu Å· Lu Hou Å· Niu Minzhe Å· Xiaodan Liang Å· Lewei Yao Å· Runhui Huang Å· Wei Zhang Å· Xin Jiang Å· Chunjing XU Å· Hang Xu
- abstract@[open-review](#): Vision-Language Pre-training (VLP) models have shown remarkable performance on various downstream tasks. Their success heavily relies on the scale of pre-trained cross-modal datasets. However, the lack of large-scale datasets and benchmarks in Chinese hinders the development of Chinese VLP models and broader multilingual applications. In this work, we release a large-scale Chinese cross-modal dataset named Wukong, which contains 100 million Chinese image-text pairs collected from the web. Wukong aims to benchmark different multi-modal pre-training methods to facilitate the VLP research and community development. Furthermore, we release a group of models pre-trained with various image encoders (ViT-B/ViT-L/SwinT) and also apply advanced pre-training techniques into VLP such as locked-image text tuning, token-wise similarity in contrastive learning, and reduced-token interaction. Extensive experiments and a benchmarking of different downstream tasks including a new largest human-verified image-text test dataset are also provided. Experiments show that Wukong can serve as a promising Chinese pre-training dataset and benchmark for different cross-modal learning methods. For the zero-shot image classification task on 10 datasets,  $\$Wukong\_text\{ViT-L\}$  achieves an average accuracy of 73.03%. For the image-text retrieval task, it achieves a mean recall of 71.6% on AIC-ICC which is 12.9% higher than WenLan 2.0. Also, our Wukong models are benchmarked on downstream tasks with other variants on multiple datasets, e.g., Flickr8K-CN, Flickr-30K-CN, COCO-CN, et al. More information can be referred to <https://wukong-dataset.github.io/wukong-dataset/>.

## [TGEA 2.0: A Large-Scale Diagnostically Annotated Dataset with Benchmark Tasks for Text Generation of Pretrained Language Models](#)

- Huibin Ge Å· Xiaohu Zhao Å· Chuang Liu Å· Yulong Zeng Å· Qun Liu Å· Deyi Xiong
- abstract@[open-review](#): In order to diagnostically analyze and improve the capability of pretrained language models (PLMs) in text generation, we propose TGEA 2.0, to date the largest dataset built on machine-authored texts by PLMs with fine-grained semantic annotations on a wide variety of pathological generation errors. We collect 170K nominal, phrasal and sentential prompts from 6M natural sentences in 3 domains. These prompts are fed into 4 generative PLMs with their best decoding strategy to generate paragraphs. 195,629 sentences are extracted from these generated paragraphs for manual annotation, where 36K erroneous sentences are detected, 42K erroneous spans are located and categorized into an error type defined in a two-level error taxonomy. We define a \textbf{M}i\textbf{S}emantic \textbf{E}rror \textbf{W}ords (MiSEW) for each erroneous span, which not only provides error-associated words but also rationalizes the reasoning behind the error. Quality control with a pre-annotation and feedback loop is performed before and during the entire annotation process. With the diagnostically annotated dataset, we propose 5 diagnosis benchmark tasks (i.e., erroneous text detection, MiSEW extraction, erroneous span location and correction together with error type classification) and 2 pathology mitigation benchmark tasks (pairwise comparison and word prediction). Experiment results on these benchmark tasks demonstrate that TGEA 2.0 is a challenging dataset that could facilitate further research on automatic diagnosis and pathology mitigation over machine texts. The dataset will be publicly available at <https://github.com/tjunlp-lab/TGEA/>.

## [BackdoorBench: A Comprehensive Benchmark of Backdoor Learning](#)

- Baoyuan Wu Å· Hongrui Chen Å· Mingda Zhang Å· Zihao Zhu Å· Shaokui Wei Å· Danni Yuan Å· Chao Shen
- abstract@[open-review](#): Backdoor learning is an emerging and vital topic for studying deep neural networks' vulnerability (DNNs). Many pioneering backdoor attack and defense methods are being proposed, successively or concurrently, in the status of a rapid arms race. However, we find that the evaluations of new methods are often unthorough to verify their claims and accurate performance, mainly due to the rapid development, diverse settings, and the difficulties of implementation and reproducibility. Without thorough evaluations and comparisons, it is not easy to track the current progress and design the future development roadmap of the literature. To alleviate this dilemma, we build a comprehensive benchmark of backdoor learning called BackdoorBench. It consists of an extensible modular-based codebase (currently including implementations of 8 state-of-the-art (SOTA) attacks and 9 SOTA defense algorithms) and a standardized protocol of complete backdoor learning. We also provide comprehensive evaluations of every pair of 8 attacks against 9 defenses, with 5 poisoning ratios, based on 5 models and 4 datasets, thus 8,000 pairs of evaluations in total. We present abundant analysis from different perspectives about these 8,000 evaluations, studying the effects of different factors in backdoor learning. All codes and evaluations of BackdoorBench are publicly available at <https://backdoorbench.github.io>.

## [ADBench: Anomaly Detection Benchmark](#)

- Songqiao Han Å· Xiyang Hu Å· Hailiang Huang Å· Minqi Jiang Å· Yue Zhao
- abstract@[open-review](#): Given a long list of anomaly detection algorithms developed in the last few decades, how do they perform with regard to (i) varying levels of supervision, (ii) different types of anomalies, and (iii) noisy and corrupted data? In this work, we answer these key questions by conducting (to our best knowledge) the most comprehensive anomaly detection benchmark with 30 algorithms on 57 benchmark datasets, named ADBench. Our extensive experiments (98,436 in total) identify meaningful insights into the role of supervision and anomaly types, and unlock future directions for researchers in algorithm selection and design. With ADBench, researchers can easily conduct comprehensive and fair evaluations for newly proposed methods on the datasets (including our contributed ones from natural language and computer vision domains) against the existing baselines. To foster accessibility and reproducibility, we fully open-source ADBench and the corresponding results.

## [M4Singer: A Multi-Style, Multi-Singer and Musical Score Provided Mandarin Singing Corpus](#)

- Lichao Zhang Å· Ruiqi Li Å· Shoutong Wang Å· Liqun Deng Å· Jinglin Liu Å· Yi Ren Å· Jinzheng He Å· Rongjie Huang Å· Jieming Zhu Å· Xiao Chen Å· Zhou Zhao
- abstract@[open-review](#): The lack of publicly available high-quality and accurately labeled datasets has long been a major bottleneck for singing voice synthesis (SVS). To tackle this problem, we present M4Singer, a free-to-use Multi-style, Multi-singer Mandarin singing collection with elaborately annotated Musical scores as well as its benchmarks. Specifically, 1) we construct and release a large high-quality Chinese singing voice corpus, which is recorded by 20 professional singers, covering 700 Chinese pop songs as well as all the four SATB types (i.e., soprano, alto, tenor, and bass); 2) we take

extensive efforts to manually compose the musical scores for each recorded song, which are necessary to the study of the prosody modeling for SVS. 3) To facilitate the use and demonstrate the quality of M4Singer, we conduct four different benchmark experiments: score-based SVS, controllable singing voice (CSV), singing voice conversion (SVC) and automatic music transcription (AMT).

## [IKEA-Manual: Seeing Shape Assembly Step by Step](#)

- Ruocheng Wang · Yunzhi Zhang · Jiayuan Mao · Ran Zhang · Chin-Yi Cheng · Jiajun Wu
- abstract@[open-review](#): Human-designed visual manuals are crucial components in shape assembly activities. They provide step-by-step guidance on how we should move and connect different parts in a convenient and physically-realizable way. While there has been an ongoing effort in building agents that perform assembly tasks, the information in human-design manuals has been largely overlooked. We identify that this is due to 1) a lack of realistic 3D assembly objects that have paired manuals and 2) the difficulty of extracting structured information from purely image-based manuals. Motivated by this observation, we present IKEA-Manual, a dataset consisting of 102 IKEA objects paired with assembly manuals. We provide fine-grained annotations on the IKEA objects and assembly manuals, including decomposed assembly parts, assembly plans, manual segmentation, and 2D-3D correspondence between 3D parts and visual manuals. We illustrate the broad application of our dataset on four tasks related to shape assembly: assembly plan generation, part segmentation, pose estimation and 3D part assembly.

## [GriddlyJS: A Web IDE for Reinforcement Learning](#)

- Christopher Bamford · Minqi Jiang · Mikayel Samvelyan · Tim Rocktäschel
- abstract@[open-review](#): Progress in reinforcement learning (RL) research is often driven by the design of new, challenging environments---a costly undertaking requiring skills orthogonal to that of a typical machine learning researcher. The complexity of environment development has only increased with the rise of procedural-content generation (PCG) as the prevailing paradigm for producing varied environments capable of testing the robustness and generalization of RL agents. Moreover, existing environments often require complex build processes, making reproducing results difficult. To address these issues, we introduce GriddlyJS, a web-based Integrated Development Environment (IDE) based on the Griddly engine. GriddlyJS allows researchers to easily design and debug arbitrary, complex PCG grid-world environments, as well as visualize, evaluate, and record the performance of trained agent models. By connecting the RL workflow to the advanced functionality enabled by modern web standards, GriddlyJS allows publishing interactive agent-environment demos that reproduce experimental results directly to the web. To demonstrate the versatility of GriddlyJS, we use it to quickly develop a complex compositional puzzle-solving environment alongside arbitrary human-designed environment configurations and their solutions for use in a automatic curriculum learning and offline RL context. The GriddlyJS IDE is open source and freely available at <https://griddly.ai>.

## [Ambiguous Images With Human Judgments for Robust Visual Event Classification](#)

- Kate Sanders · Reno Kriz · Anqi Liu · Benjamin Van Durme
- abstract@[open-review](#): Contemporary vision benchmarks predominantly consider tasks on which humans can achieve near-perfect performance. However, humans are frequently presented with visual data that they cannot classify with 100% certainty, and models trained on standard vision benchmarks achieve low performance when evaluated on this data. To address this issue, we introduce a procedure for creating datasets of ambiguous images and use it to produce SQUID-E ("Squidy"), a collection of noisy images extracted from videos. All images are annotated with ground truth values and a test set is annotated with human uncertainty judgments. We use this dataset to characterize human uncertainty in vision tasks and evaluate existing visual event classification models. Experimental results suggest that existing vision models are not sufficiently equipped to provide meaningful outputs for ambiguous images and that datasets of this nature can be used to assess and improve such models through model training and direct evaluation of model calibration. These findings motivate large-scale ambiguous dataset creation and further research focusing on noisy visual data.

## [EnvPool: A Highly Parallel Reinforcement Learning Environment Execution Engine](#)

- Jiayi Weng · Min Lin · Shengyi Huang · Bo Liu · Denys Makoviichuk · Viktor Makoviychuk · Zichen Liu · Yufan Song · Ting Luo · Yukun Jiang · Zhongwen Xu · Shuicheng Yan
- abstract@[open-review](#): There has been significant progress in developing reinforcement learning (RL) training systems. Past works such as IMPALA, Apex, Seed RL, Sample Factory, and others, aim to improve the system's overall throughput. In this paper, we aim to address a common bottleneck in the RL training system, i.e., parallel environment execution, which is often the slowest part of the whole system but receives little attention. With a curated design for paralleling RL environments, we have improved the RL environment simulation speed across different hardware setups, ranging from a laptop and a modest workstation, to a high-end machine such as NVIDIA DGX-A100. On a high-end machine, EnvPool achieves one million frames per second for the environment execution on Atari environments and three million frames per second on MuJoCo environments. When running EnvPool on a laptop, the speed is 2.8x that of the Python subprocess. Moreover, great compatibility with existing RL training libraries has been demonstrated in the open-sourced community, including CleanRL, rl\_games, DeepMind Acme, etc. Finally, EnvPool allows researchers to iterate their ideas at a much faster pace and has great potential to become the de facto RL environment execution engine. Example runs show that it only takes five minutes to train agents to play Atari Pong and MuJoCo Ant on a laptop. EnvPool is open-sourced at <https://github.com/sail-sg/envpool>.

## [Evaluating Out-of-Distribution Performance on Document Image Classifiers](#)

- Stefan Larson · Yi Yang · Gordon Lim · Yutong Ai · David Kuang · Kevin Leach
- abstract@[open-review](#): The ability of a document classifier to handle inputs that are drawn from a distribution different from the training distribution is crucial for robust deployment and generalizability. The RVL-CDIP corpus is the de facto standard benchmark for document classification, yet to our knowledge all studies that use this corpus do not include evaluation on out-of-distribution documents. In this paper, we curate and release a new out-of-distribution benchmark for evaluating out-of-distribution performance for document classifiers. Our new out-of-distribution benchmark consists of two types of documents: those that are not part of any of the 16 in-domain RVL-CDIP categories (RVL-CDIP-N), and those that are one of the 16 in-domain categories yet are drawn from a distribution different from that of the original RVL-CDIP dataset (RVL-CDIP-O). While prior work on document classification for in-domain RVL-CDIP documents reports high accuracy scores, we find that these models exhibit accuracy drops of between roughly 15-30% on our new out-of-domain RVL-CDIP-N benchmark. Our new benchmark researchers with a valuable new resource for analyzing out-of-distribution performance on document classifiers.

## [USB: A Unified Semi-supervised Learning Benchmark for Classification](#)

- Yidong Wang · Hao Chen · Yue Fan · Wang SUN · Ran Tao · Wenxin Hou · Renjie Wang · Linyi Yang · Zhi Zhou · Lan-Zhe Guo · Heli Qi · Zhen Wu · Yu-Feng Li · Satoshi Nakamura · Wei Ye · Marios Savvides · Bhiksha Raj · Takahiro Shinozaki · Bernt Schiele · Jindong Wang · Xing Xie · Yue Zhang
- abstract@[open-review](#): Semi-supervised learning (SSL) improves model generalization by leveraging massive unlabeled data to augment limited labeled samples. However, currently, popular SSL evaluation protocols are often constrained to computer vision (CV) tasks. In addition, previous work typically trains deep neural networks from scratch, which is time-consuming and environmentally unfriendly. To address the above issues, we construct a Unified SSL Benchmark (USB) for classification by selecting 15 diverse, challenging, and comprehensive tasks from CV, natural language processing (NLP), and audio processing (Audio), on which we systematically evaluate the dominant SSL methods, and also open-source a modular and extensible codebase for fair evaluation of these SSL methods. We further provide the pre-trained versions of the state-of-the-art neural models for CV tasks to make the cost affordable for further tuning. USB enables the evaluation of a single SSL algorithm on more tasks from multiple domains but with less cost. Specifically,

on a single NVIDIA V100, only 39 GPU days are required to evaluate FixMatch on 15 tasks in USB while 335 GPU days (279 GPU days on 4 CV datasets except for ImageNet) are needed on 5 CV tasks with TorchSSL.

## [OpenSRH: optimizing brain tumor surgery using intraoperative stimulated Raman histology](#)

- Cheng Jiang · Asadur Chowdury · Xinhai Hou · Akhil Kondepudi · Christian Freudiger · Kyle Conway · Sandra Camelo-Piragua · Daniel Orringer · Honglak Lee · Todd Hollon
- abstract@[open-review](#): Accurate intraoperative diagnosis is essential for providing safe and effective care during brain tumor surgery. Our standard-of-care diagnostic methods are time, resource, and labor intensive, which restricts access to optimal surgical treatments. To address these limitations, we propose an alternative workflow that combines stimulated Raman histology (SRH), a rapid optical imaging method, with deep learning-based automated interpretation of SRH images for intraoperative brain tumor diagnosis and real-time surgical decision support. Here, we present OpenSRH, the first public dataset of clinical SRH images from 300+ brain tumors patients and 1300+ unique whole slide optical images. OpenSRH contains data from the most common brain tumors diagnoses, full pathologic annotations, whole slide tumor segmentations, raw and processed optical imaging data for end-to-end model development and validation. We provide a framework for patch-based whole slide SRH classification and inference using weak (i.e. patient-level) diagnostic labels. Finally, we benchmark two computer vision tasks: multi-class histologic brain tumor classification and patch-based contrastive representation learning. We hope OpenSRH will facilitate the clinical translation of rapid optical imaging and real-time ML-based surgical decision support in order to improve the access, safety, and efficacy of cancer surgery in the era of precision medicine.

## [Video compression dataset and benchmark of learning-based video-quality metrics](#)

- Anastasia Antsiferova · Sergey Lavrushkin · Maksim Smirnov · Aleksandr Gushchin · Dmitriy Vatolin · Dmitriy Kulikov
- abstract@[open-review](#): Video-quality measurement is a critical task in video processing. Nowadays, many implementations of new encoding standards - such as AV1, VVC, and LCEVC - use deep-learning-based decoding algorithms with perceptual metrics that serve as optimization objectives. But investigations of the performance of modern video- and image-quality metrics commonly employ videos compressed using older standards, such as AVC. In this paper, we present a new benchmark for video-quality metrics that evaluates video compression. It is based on a new dataset consisting of about 2,500 streams encoded using different standards, including AVC, HEVC, AV1, VP9, and VVC. Subjective scores were collected using crowdsourced pairwise comparisons. The list of evaluated metrics includes recent ones based on machine learning and neural networks. The results demonstrate that new no-reference metrics exhibit high correlation with subjective quality and approach the capability of top full-reference metrics.

## [Flare7K: A Phenomenological Nighttime Flare Removal Dataset](#)

- Yuekun Dai · Chongyi Li · Shangchen Zhou · Ruicheng Feng · Chen Change Loy
- abstract@[open-review](#): Artificial lights commonly leave strong lens flare artifacts on images captured at night. Nighttime flare not only affects the visual quality but also degrades the performance of vision algorithms. Existing flare removal methods mainly focus on removing daytime flares and fail in nighttime. Nighttime flare removal is challenging because of the unique luminance and spectrum of artificial lights and the diverse patterns and image degradation of the flares captured at night. The scarcity of nighttime flare removal datasets limits the research on this crucial task. In this paper, we introduce, Flare7K, the first nighttime flare removal dataset, which is generated based on the observation and statistics of real-world nighttime lens flares. It offers 5,000 scattering and 2,000 reflective flare images, consisting of 25 types of scattering flares and 10 types of reflective flares. The 7,000 flare patterns can be randomly added to flare-free images, forming the flare-corrupted and flare-free image pairs. With the paired data, we can train deep models to restore flare-corrupted images taken in the real world effectively. Apart from abundant flare patterns, we also provide rich annotations, including the labeling of light source, glare with shimmer, reflective flare, and streak, which are commonly absent from existing datasets. Hence, our dataset can facilitate new work in nighttime flare removal and more fine-grained analysis of flare patterns. Extensive experiments show that our dataset adds diversity to existing flare datasets and pushes the frontier of nighttime flare removal.

## [GOOD: A Graph Out-of-Distribution Benchmark](#)

- Shurui Gui · Xiner Li · Limei Wang · Shuiwang Ji
- abstract@[open-review](#): Out-of-distribution (OOD) learning deals with scenarios in which training and test data follow different distributions. Although general OOD problems have been intensively studied in machine learning, graph OOD is only an emerging area of research. Currently, there lacks a systematic benchmark tailored to graph OOD method evaluation. In this work, we aim at developing an OOD benchmark, known as GOOD, for graphs specifically. We explicitly make distinctions between covariate and concept shifts and design data splits that accurately reflect different shifts. We consider both graph and node prediction tasks as there are key differences in designing shifts. Overall, GOOD contains 11 datasets with 17 domain selections. When combined with covariate, concept, and no shifts, we obtain 51 different splits. We provide performance results on 10 commonly used baseline methods with 10 random runs. This results in 510 dataset-model combinations in total. Our results show significant performance gaps between in-distribution and OOD settings. Our results also shed light on different performance trends between covariate and concept shifts by different methods. Our GOOD benchmark is a growing project and expects to expand in both quantity and variety of resources as the area develops. The GOOD benchmark can be accessed via <https://github.com/divelab/GOOD/>.

## [A new dataset for multilingual keyphrase generation](#)

- FrÃ©dÃ©ric Piedboeuf · Philippe Langlais
- abstract@[open-review](#): Keyphrases are an important tool for efficiently dealing with the ever-increasing amount of information present on the internet. While there are many recent papers on English keyphrase generation, keyphrase generation for other languages remains vastly understudied, mostly due to the absence of datasets. To address this, we present a novel dataset called Papyrus, composed of 16427 pairs of abstracts and keyphrases. We release four versions of this dataset, corresponding to different subtasks. Papyrus-e considers only English keyphrases, Papyrus-f considers French keyphrases, Papyrus-m considers keyphrase generation in any language (mostly French and English), and Papyrus-a considers keyphrase generation in several languages. We train a state-of-the-art model on all four tasks and show that they lead to better results for non-English languages, with an average improvement of 14.2% on keyphrase extraction and 2.0% on generation. We also show an improvement of 0.4% on extraction and 0.7% on generation over English state-of-the-art results by concatenating Papyrus-e with the Kp20K training set.

## [TaiSu: A 166M Large-scale High-Quality Dataset for Chinese Vision-Language Pre-training](#)

- Yulong Liu · Guibo Zhu · Bin Zhu · Qi Song · Guojing Ge · Haoran Chen · GuanHui Qiao · Ru Peng · Lingxiang Wu · Jinqiao Wang
- abstract@[open-review](#): Vision-Language Pre-training (VLP) has been shown to be an efficient method to improve the performance of models on different vision-and-language downstream tasks. Substantial studies have shown that neural networks may be able to learn some general rules about language and visual concepts from a large-scale weakly labeled image-text dataset. However, most of the public cross-modal datasets that contain more than 100M image-text pairs are in English; there is a lack of available large-scale and high-quality Chinese VLP datasets. In this work, we propose a new framework for automatic dataset acquisition and cleaning with which we construct a new large-scale and high-quality cross-modal dataset named as TaiSu, containing 166 million images and 219 million Chinese captions. Compared with the recently released Wukong dataset, our dataset is achieved with much stricter restrictions on the semantic correlation of image-text pairs. We also propose to combine texts collected from the web with texts generated by a pre-trained image-captioning model. To the best of our knowledge, TaiSu is currently the largest publicly accessible Chinese cross-modal dataset. Furthermore, we test our dataset on several vision-language downstream tasks. TaiSu outperforms BriVL by a large margin on the zero-shot image-text retrieval task and

zero-shot image classification task. TaiSu also shows better performance than Wukong on the image-retrieval task without using image augmentation for training. Results demonstrate that TaiSu can serve as a promising VLP dataset, both for understanding and generative tasks. More information can be referred to <https://github.com/ksOAn6g5/TaiSu>.

## [VLMbench: A Compositional Benchmark for Vision-and-Language Manipulation](#)

- Kaizhi Zheng · Xiaotong Chen · Odest Chadwicke Jenkins · Xin Wang
- abstract@[open-review](#): Benefiting from language flexibility and compositionality, humans naturally intend to use language to command an embodied agent for complex tasks such as navigation and object manipulation. In this work, we aim to fill the blank of the last mile of embodied agents---object manipulation by following human guidance, e.g., “move the red mug next to the box while keeping it upright.” To this end, we introduce an Automatic Manipulation Solver (AMSolver) system and build a Vision-and-Language Manipulation benchmark (VLMbench) based on it, containing various language instructions on categorized robotic manipulation tasks. Specifically, modular rule-based task templates are created to automatically generate robot demonstrations with language instructions, consisting of diverse object shapes and appearances, action types, and motion constraints. We also develop a keypoint-based model 6D-CLIPort to deal with multi-view observations and language input and output a sequence of 6 degrees of freedom (DoF) actions. We hope the new simulator and benchmark will facilitate future research on language-guided robotic manipulation.

## [ViSioNS: Visual Search in Natural Scenes Benchmark](#)

- Fermán Travi · Gonzalo Ruarte · Gastón Bujia · Juan Esteban Kamienkowski
- abstract@[open-review](#): Visual search is an essential part of almost any everyday human interaction with the visual environment. Nowadays, several algorithms are able to predict gaze positions during simple observation, but few models attempt to simulate human behavior during visual search in natural scenes. Furthermore, these models vary widely in their design and exhibit differences in the datasets and metrics with which they were evaluated. Thus, there is a need for a reference point, on which each model can be tested and from where potential improvements can be derived. In this study, we select publicly available state-of-the-art visual search models and datasets in natural scenes, and provide a common framework for their evaluation. To this end, we apply a unified format and criteria, bridging the gaps between them, and we estimate the models’ efficiency and similarity with humans using a specific set of metrics. This integration has allowed us to enhance the Ideal Bayesian Searcher by combining it with a neural network-based visual search model, which enables it to generalize to other datasets. The present work sheds light on the limitations of current models and how integrating different approaches with a unified criteria can lead to better algorithms. Moreover, it moves forward on bringing forth a solution for the urgent need for benchmarking data and metrics to support the development of more general human visual search computational models. All of the code used here, including metrics, plots, and visual search models, alongside the preprocessed datasets, are available at [\\\$url{https://github.com/FerminT/VisualSearchBenchmark} \\$](https://github.com/FerminT/VisualSearchBenchmark).

## [Ontologue: Declarative Benchmark Construction for Ontological Multi-Label Classification](#)

- Sean Yang · Bernease Herman · Bill Howe
- abstract@[open-review](#): We describe a customizable benchmark for hierarchical and ontological multi-label classification, a task where labels are equipped with a graph structure and data items can be assigned multiple labels. We find that current benchmarks do not adequately represent the problem space, casting doubt on the generalizability of current results. We consider three dimensions of the problem space: context (availability of rich features on the data and labels), distribution of labels over data, and graph structure. For context, the lack of complex features on the labels (and in some cases, the data) artificially prevent the use of modern representation learning techniques as an appropriate baseline. For distribution, we find the long tail of labels over data constitute a few-shot learning problem that artificially confounds the results: for most common benchmarks, over 40% of the labels have fewer than 5 data points in the training set. For structure, we find that the correlation between performance and the height of the tree can explain some of the variation in performance, informing practical utility. In this paper, we demonstrate how the lack of diversity in benchmarks can confound performance analysis, then present a declarative query system called Ontologue for generating custom benchmarks with specific properties, then use this system to design 4 new benchmarks extracted from DBpedia that better represent the problem space. We evaluate state-of-the-art algorithms on both existing and new benchmarks and show that the performance conclusions can vary significantly depending on the dimensions we consider. We intend the system and derived benchmarks to improve the analysis of generalizability for these problems.

## [WinoGAViL: Gamified Association Benchmark to Challenge Vision-and-Language Models](#)

- Yonatan Bitton · Nitzan Bitton Guetta · Ron Yosef · Yuval Elovici · Mohit Bansal · Gabriel Stanovsky · Roy Schwartz
- abstract@[open-review](#): While vision-and-language models perform well on tasks such as visual question answering, they struggle when it comes to basic human commonsense reasoning skills. In this work, we introduce WinoGAViL: an online game of vision-and-language associations (e.g., between werewolves and a full moon), used as a dynamic evaluation benchmark. Inspired by the popular card game Codenames, a spymaster gives a textual cue related to several visual candidates, and another player tries to identify them. Human players are rewarded for creating associations that are challenging for a rival AI model but still solvable by other human players. We use the game to collect 3.5K instances, finding that they are intuitive for humans (>90% Jaccard index) but challenging for state-of-the-art AI models, where the best model (ViLT) achieves a score of 52%, succeeding mostly where the cue is visually salient. Our analysis as well as the feedback we collect from players indicate that the collected associations require diverse reasoning skills, including general knowledge, common sense, abstraction, and more. We release the dataset, the code and the interactive game, allowing future data collection that can be used to develop models with better association abilities.

## [Multi-LexSum: Real-world Summaries of Civil Rights Lawsuits at Multiple Granularities](#)

- Zejiang Shen · Kyle Lo · Lauren Yu · Nathan Dahlberg · Margo Schlanger · Doug Downey
- abstract@[open-review](#): With the advent of large language models, methods for abstractive summarization have made great strides, creating potential for use in applications to aid knowledge workers processing unwieldy document collections. One such setting is the Civil Rights Litigation Clearinghouse (CRLC, <https://clearinghouse.net>), which posts information about large-scale civil rights lawsuits, serving lawyers, scholars, and the general public. Today, summarization in the CRLC requires extensive training of lawyers and law students who spend hours per case understanding multiple relevant documents in order to produce high-quality summaries of key events and outcomes. Motivated by this ongoing real-world summarization effort, we introduce Multi-LexSum, a collection of 9,280 expert-authored summaries drawn from ongoing CRLC writing. Multi-LexSum presents a challenging multi-document summarization task given the length of the source documents, often exceeding two hundred pages per case. Furthermore, Multi-LexSum is distinct from other datasets in its multiple target summaries, each at a different granularity (ranging from one-sentence “extreme” summaries to multi-paragraph narrations of over five hundred words). We present extensive analysis demonstrating that despite the high-quality summaries in the training data (adhering to strict content and style guidelines), state-of-the-art summarization models perform poorly on this task. We release Multi-LexSum for further summarization research and to facilitate the development of applications to assist in the CRLC’s mission at <https://multilexsum.github.io>.

## [OpenFilter: A Framework to Democratize Research Access to Social Media AR Filters](#)

- Piera Riccio · Bill Psomas · Francesco Galati · Francisco Escolano · Thomas Hofmann · Nuria Oliver
- abstract@[open-review](#): Augmented Reality or AR filters on selfies have become very popular on social media platforms for a variety of applications, including marketing, entertainment and aesthetics. Given the wide adoption of AR face filters and the importance of faces in our social structures and relations, there is increased interest by the scientific community to analyze the impact of such filters from a psychological, artistic and sociological

perspective. However, there are few quantitative analyses in this area mainly due to a lack of publicly available datasets of facial images with applied AR filters. The proprietary, close nature of most social media platforms does not allow users, scientists and practitioners to access the code and the details of the available AR face filters. Scraping faces from these platforms to collect data is ethically unacceptable and should, therefore, be avoided in research. In this paper, we present OpenFilter, a flexible framework to apply AR filters available in social media platforms on existing large collections of human faces. Moreover, we share FairBeauty and B-LFW, two beautified versions of the publicly available FairFace and LFW datasets and we outline insights derived from the analysis of these beautified datasets.

## [Towards Video Text Visual Question Answering: Benchmark and Baseline](#)

- Minyi Zhao · Bingjia Li · Jie Wang · Wanqing Li · Wenjing Zhou · Lan Zhang · Shijie Xuyang · Zhihang Yu · Xinkun Yu · Guangze Li · Aobotao Dai · Shuigeng Zhou
- abstract@[open-review](#): There are already some text-based visual question answering (TextVQA) benchmarks for developing machine's ability to answer questions based on texts in images in recent years. However, models developed on these benchmarks cannot work effectively in many real-life scenarios (e.g. traffic monitoring, shopping ads and e-learning videos) where temporal reasoning ability is required. To this end, we propose a new task named Video Text Visual Question Answering (ViteVQA in short) that aims at answering questions by reasoning texts and visual information spatiotemporally in a given video. In particular, on the one hand, we build the first ViteVQA benchmark dataset named M4-ViteVQA --- the abbreviation of Multi-category Multi-frame Multi-resolution Multi-modal benchmark for ViteVQA, which contains 7,620 video clips of 9 categories (i.e., shopping, traveling, driving, vlog, sport, advertisement, movie, game and talking) and 3 kinds of resolutions (i.e., 720p, 1080p and 1176x664), and 25,123 question-answer pairs. On the other hand, we develop a baseline method named T5-ViteVQA for the ViteVQA task. T5-ViteVQA consists of five transformers. It first extracts optical character recognition (OCR) tokens, question features, and video representations via two OCR transformers, one language transformer and one video-language transformer, respectively. Then, a multimodal fusion transformer and an answer generation module are applied to fuse multimodal information and generate the final prediction. Extensive experiments on M4-ViteVQA demonstrate the superiority of T5-ViteVQA to the existing approaches of TextVQA and VQA tasks. The ViteVQA benchmark is available in <https://github.com/bytedance/VTVQA>.

## [Breaking Bad: A Dataset for Geometric Fracture and Reassembly](#)

- Silvia Sellßen · Yun-Chun Chen · Ziyi Wu · Animesh Garg · Alec Jacobson
- abstract@[open-review](#): We introduce Breaking Bad, a large-scale dataset of fractured objects. Our dataset consists of over one million fractured objects simulated from ten thousand base models. The fracture simulation is powered by a recent physically based algorithm that efficiently generates a variety of fracture modes of an object. Existing shape assembly datasets decompose objects according to semantically meaningful parts, effectively modeling the construction process. In contrast, Breaking Bad models the destruction process of how a geometric object naturally breaks into fragments. Our dataset serves as a benchmark that enables the study of fractured object reassembly and presents new challenges for geometric shape understanding. We analyze our dataset with several geometry measurements and benchmark three state-of-the-art shape assembly deep learning methods under various settings. Extensive experimental results demonstrate the difficulty of our dataset, calling on future research in model designs specifically for the geometric shape assembly task. We host our dataset at <https://breaking-bad-dataset.github.io/>.

## [Model Zoos: A Dataset of Diverse Populations of Neural Network Models](#)

- Konstantin Schärlholt · Diyar Taskiran · Boris Knyazev · Xavier Giro-i-Nieto · Damian Borth
- abstract@[open-review](#): In the last years, neural networks (NN) have evolved from laboratory environments to the state-of-the-art for many real-world problems. It was shown that NN models (i.e., their weights and biases) evolve on unique trajectories in weight space during training. Following, a population of such neural network models (referred to as model zoo) would form structures in weight space. We think that the geometry, curvature and smoothness of these structures contain information about the state of training and can reveal latent properties of individual models. With such model zoos, one could investigate novel approaches for (i) model analysis, (ii) discover unknown learning dynamics, (iii) learn rich representations of such populations, or (iv) exploit the model zoos for generative modelling of NN weights and biases. Unfortunately, the lack of standardized model zoos and available benchmarks significantly increases the friction for further research about populations of NNs. With this work, we publish a novel dataset of model zoos containing systematically generated and diverse populations of NN models for further research. In total the proposed model zoo dataset is based on eight image datasets, consists of 27 model zoos trained with varying hyperparameter combinations and includes 50<sup>TM</sup> 360 unique NN models as well as their sparsified twins, resulting in over 3<sup>TM</sup> 844<sup>TM</sup> 360 collected model states. Additionally, to the model zoo data we provide an in-depth analysis of the zoos and provide benchmarks for multiple downstream tasks. The dataset can be found at [www.modelzoos.cc](http://www.modelzoos.cc).

## [Active-Passive SimStereo - Benchmarking the Cross-Generalization Capabilities of Deep Learning-based Stereo Methods](#)

- Laurent Jospin · Allen Antony · Lian Xu · Hamid Laga · Farid Boussaid · Mohammed Bennamoun
- abstract@[open-review](#): In stereo vision, self-similar or bland regions can make it difficult to match patches between two images. Active stereo-based methods mitigate this problem by projecting a pseudo-random pattern on the scene so that each patch of an image pair can be identified without ambiguity. However, the projected pattern significantly alters the appearance of the image. If this pattern acts as a form of adversarial noise, it could negatively impact the performance of deep learning-based methods, which are now the de-facto standard for dense stereo vision. In this paper, we propose the Active-Passive SimStereo dataset and a corresponding benchmark to evaluate the performance gap between passive and active stereo images for stereo matching algorithms. Using the proposed benchmark and an additional ablation study, we show that the feature extraction and matching modules of a selection of twenty selected deep learning-based stereo matching methods generalize to active stereo without a problem. However, the disparity refinement modules of three of the twenty architectures (ACVNet, CascadeStereo, and StereoNet) are negatively affected by the active stereo patterns due to their reliance on the appearance of the input images.

## [ENS-10: A Dataset For Post-Processing Ensemble Weather Forecasts](#)

- Saleh Ashkboos · Langwen Huang · Nikoli Dryden · Tal Ben-Nun · Peter Dueben · Lukas Gianinazzi · Luca Kummer · Torsten Hoefer
- abstract@[open-review](#): Post-processing ensemble prediction systems can improve the reliability of weather forecasting, especially for extreme event prediction. In recent years, different machine learning models have been developed to improve the quality of weather post-processing. However, these models require a comprehensive dataset of weather simulations to produce high-accuracy results, which comes at a high computational cost to generate. This paper introduces the ENS-10 dataset, consisting of ten ensemble members spanning 20 years (1998--2017). The ensemble members are generated by perturbing numerical weather simulations to capture the chaotic behavior of the Earth. To represent the three-dimensional state of the atmosphere, ENS-10 provides the most relevant atmospheric variables at 11 distinct pressure levels and the surface at \ang{0.5} resolution for forecast lead times T=0, 24, and 48 hours (two data points per week). We propose the ENS-10 prediction correction task for improving the forecast quality at a 48-hour lead time through ensemble post-processing. We provide a set of baselines and compare their skill at correcting the predictions of three important atmospheric variables. Moreover, we measure the baselines' skill at improving predictions of extreme weather events using our dataset. The ENS-10 dataset is available under the Creative Commons Attribution 4.0 International (CC BY 4.0) license.

## [NAS-Bench-360: Benchmarking Neural Architecture Search on Diverse Tasks](#)

- Renbo Tu · Nicholas Roberts · Misha Khodak · Junhong Shen · Frederic Sala · Ameet Talwalkar

- abstract@[open-review](#): Most existing neural architecture search (NAS) benchmarks and algorithms prioritize well-studied tasks, e.g. image classification on CIFAR or ImageNet. This makes the performance of NAS approaches in more diverse areas poorly understood. In this paper, we present NAS-Bench-360, a benchmark suite to evaluate methods on domains beyond those traditionally studied in architecture search, and use it to address the following question: do state-of-the-art NAS methods perform well on diverse tasks? To construct the benchmark, we curate ten tasks spanning a diverse array of application domains, dataset sizes, problem dimensionalities, and learning objectives. Each task is carefully chosen to interoperate with modern CNN-based search methods while possibly being far-afield from its original development domain. To speed up and reduce the cost of NAS research, for two of the tasks we release the precomputed performance of 15,625 architectures comprising a standard CNN search space. Experimentally, we show the need for more robust NAS evaluation of the kind NAS-Bench-360 enables by showing that several modern NAS procedures perform inconsistently across the ten tasks, with many catastrophically poor results. We also demonstrate how NAS-Bench-360 and its associated precomputed results will enable future scientific discoveries by testing whether several recent hypotheses promoted in the NAS literature hold on diverse tasks. NAS-Bench-360 is hosted at <https://nb360.ml.cmu.edu>.

## [Beyond Real-world Benchmark Datasets: An Empirical Study of Node Classification with GNNs](#)

- Seiji Maekawa · Koki Noda · Yuya Sasaki · makoto onizuka
- abstract@[open-review](#): Graph Neural Networks (GNNs) have achieved great success on a node classification task. Despite the broad interest in developing and evaluating GNNs, they have been assessed with limited benchmark datasets. As a result, the existing evaluation of GNNs lacks fine-grained analysis from various characteristics of graphs. Motivated by this, we conduct extensive experiments with a synthetic graph generator that can generate graphs having controlled characteristics for fine-grained analysis. Our empirical studies clarify the strengths and weaknesses of GNNs from four major characteristics of real-world graphs with class labels of nodes, i.e., 1) class size distributions (balanced vs. imbalanced), 2) edge connection proportions between classes (homophilic vs. heterophilic), 3) attribute values (biased vs. random), and 4) graph sizes (small vs. large). In addition, to foster future research on GNNs, we publicly release our codebase that allows users to evaluate various GNNs with various graphs. We hope this work offers interesting insights for future research.

## [FlyView: a bio-inspired optical flow truth dataset for visual navigation using panoramic stereo vision](#)

- Alix Leroy · Graham Taylor
- abstract@[open-review](#): Flying at speed through complex environments is a challenging task that has been performed successfully by insects since the Carboniferous, but which remains a challenge for robotic and autonomous systems. Insects navigate the world using optical flow sensed by their compound eyes, which they process using a deep neural network weighing just a few milligrams. Deploying an insect-inspired network architecture in computer vision could therefore enable more efficient and effective ways of estimating structure and self-motion using optical flow. Training a bio-inspired deep network to implement these tasks requires biologically relevant training, test, and validation data. To this end, we introduce FlyView, a novel bio-inspired truth dataset for visual navigation. This simulated dataset is rendered using open source 3D scenes in which the observer's position is known at every frame, and is accompanied by truth data on depth, self-motion, and motion flow. This dataset comprising 42,475 frames has several key features that are missing from existing optical flow datasets, including: (i) panoramic cameras with a monocular and binocular field of view matched to that of a fly's compound eyes; (ii) dynamically meaningful self-motion modelled on motion primitives, or the 3D trajectories of drones and flies; and (iii) complex natural and indoor environments including reflective surfaces.

## [Characteristics of Harmful Text: Towards Rigorous Benchmarking of Language Models](#)

- Maribeth Rauh · John Mellor · Jonathan Uesato · Po-Sen Huang · Johannes Welbl · Laura Weidinger · Sumanth Dathathri · Amelia Glaese · Geoffrey Irving · Iason Gabriel · William Isaac · Lisa Anne Hendricks
- abstract@[open-review](#): Large language models produce human-like text that drive a growing number of applications. However, recent literature and, increasingly, real world observations, have demonstrated that these models can generate language that is toxic, biased, untruthful or otherwise harmful. Though work to evaluate language model harms is under way, translating foresight about which harms may arise into rigorous benchmarks is not straightforward. To facilitate this translation, we outline six ways of characterizing harmful text which merit explicit consideration when designing new benchmarks. We then use these characteristics as a lens to identify trends and gaps in existing benchmarks. Finally, we apply them in a case study of the Perspective API, a toxicity classifier that is widely used in harm benchmarks. Our characteristics provide one piece of the bridge that translates between foresight and effective evaluation.

## [The Dollar Street Dataset: Images Representing the Geographic and Socioeconomic Diversity of the World](#)

- William Gaviria Rojas · Sudnya Diamos · Keertan Kini · David Kanter · Vijay Janapa Reddi · Cody Coleman
- abstract@[open-review](#): It is crucial that image datasets for computer vision are representative and contain accurate demographic information to ensure their robustness and fairness, especially for smaller subpopulations. To address this issue, we present Dollar Street - a supervised dataset that contains 38,479 images of everyday household items from homes around the world. This dataset was manually curated and fully labeled, including tags for objects (e.g. toilet, toothbrush, stove) and demographic data such as region, country and home monthly income. This dataset includes images from homes with no internet access and incomes as low as \$26.99 per month, visually capturing valuable socioeconomic diversity of traditionally under-represented populations. All images and data are licensed under CC-BY, permitting their use in academic and commercial work. Moreover, we show that this dataset can improve the performance of classification tasks for images of household items from lower income homes, addressing a critical need for datasets that combat bias.

## [Pythae: Unifying Generative Autoencoders in Python - A Benchmarking Use Case](#)

- Clément Chadebec · Louis Vincent · Stephanie Allassonnière
- abstract@[open-review](#): In recent years, deep generative models have attracted increasing interest due to their capacity to model complex distributions. Among those models, variational autoencoders have gained popularity as they have proven both to be computationally efficient and yield impressive results in multiple fields. Following this breakthrough, extensive research has been done in order to improve the original publication, resulting in a variety of different VAE models in response to different tasks. In this paper we present `\textbf{Pythae}`, a versatile `\textit{open-source}` Python library providing both a `\textit{unified implementation}` and a dedicated framework allowing `\textit{straightforward}`, `\textit{reproducible}` and `\textit{reliable}` use of generative autoencoder models. We then propose to use this library to perform a case study benchmark where we present and compare 19 generative autoencoder models representative of some of the main improvements on downstream tasks such as image reconstruction, generation, classification, clustering and interpolation. The open-source library can be found at `\url{https://github.com/clementchadebec/benchmark_VAE}`.

## [Robustness Analysis of Video-Language Models Against Visual and Language Perturbations](#)

- Madeline Chantry · Shruti Vyas · Hamid Palangi · Yogesh Rawat · Vibhav Vineet
- abstract@[open-review](#): Joint visual and language modeling on large-scale datasets has recently shown good progress in multi-modal tasks when compared to single modal learning. However, robustness of these approaches against real-world perturbations has not been studied. In this work, we perform the first extensive robustness study of video-language models against various real-world perturbations. We focus on text-to-video retrieval and propose two large-scale benchmark datasets, MSRVTT-P and YouCook2-P, which utilize 90 different visual and 35 different text perturbations. The study reveals some interesting initial findings from the studied models: 1) models are more robust when text is perturbed versus when video is perturbed, 2) models that are

pre-trained are more robust than those trained from scratch, 3) models attend more to scene and objects rather than motion and action. We hope this study will serve as a benchmark and guide future research in robust video-language learning. The benchmark introduced in this study along with the code and datasets is available at <https://bit.ly/3CNOly4>.

## [Benchmarking Heterogeneous Treatment Effect Models through the Lens of Interpretability](#)

- Jonathan CrabbÃ© · Alicia Curth · Ioana Bica · Mihaela van der Schaar
- abstract@[open-review](#): Estimating personalized effects of treatments is a complex, yet pervasive problem. To tackle it, recent developments in the machine learning (ML) literature on heterogeneous treatment effect estimation gave rise to many sophisticated, but opaque, tools: due to their flexibility, modularity and ability to learn constrained representations, neural networks in particular have become central to this literature. Unfortunately, the assets of such black boxes come at a cost: models typically involve countless nontrivial operations, making it difficult to understand what they have learned. Yet, understanding these models can be crucial -- in a medical context, for example, discovered knowledge on treatment effect heterogeneity could inform treatment prescription in clinical practice. In this work, we therefore use post-hoc feature importance methods to identify features that influence the model's predictions. This allows us to evaluate treatment effect estimators along a new and important dimension that has been overlooked in previous work: We construct a benchmarking environment to empirically investigate the ability of personalized treatment effect models to identify predictive covariates -- covariates that determine differential responses to treatment. Our benchmarking environment then enables us to provide new insight into the strengths and weaknesses of different types of treatment effects models as we modulate different challenges specific to treatment effect estimation -- e.g. the ratio of prognostic to predictive information, the possible nonlinearity of potential outcomes and the presence and type of confounding.

## [A Comprehensive Study on Large-Scale Graph Training: Benchmarking and Rethinking](#)

- Keyu Duan · Zirui Liu · Peihao Wang · Wenqing Zheng · Kaixiong Zhou · Tianlong Chen · Xia Hu · Zhangyang Wang
- abstract@[open-review](#): Large-scale graph training is a notoriously challenging problem for graph neural networks (GNNs). Due to the nature of evolving graph structures into the training process, vanilla GNNs usually fail to scale up, limited by the GPU memory space. Up to now, though numerous scalable GNN architectures have been proposed, we still lack a comprehensive survey and fair benchmark of this reservoir to find the rationale for designing scalable GNNs. To this end, we first systematically formulate the representative methods of large-scale graph training into several branches and further establish a fair and consistent benchmark for them by a greedy hyperparameter searching. In addition, regarding efficiency, we theoretically evaluate the time and space complexity of various branches and empirically compare them w.r.t GPU memory usage, throughput, and convergence. Furthermore, We analyze the pros and cons for various branches of scalable GNNs and then present a new ensembling training manner, named EnGCN, to address the existing issues. Remarkably, our proposed method has achieved new state-of-the-art (SOTA) performance on large-scale datasets. Our code is available at [https://github.com/VITA-Group/LargeScaleGCN\\_Benchmarking](https://github.com/VITA-Group/LargeScaleGCN_Benchmarking).

## [FETA: Towards Specializing Foundational Models for Expert Task Applications](#)

- Amit Alfassy · Assaf Arbelle · Oshri Halimi · Sivan Harary · Roei Herzig · Eli Schwartz · Rameswar Panda · Michele Dolfi · Christoph Auer · Peter Staar · Kate Saenko · Rogerio Feris · Leonid Karlinsky
- abstract@[open-review](#): Foundational Models (FMs) have demonstrated unprecedented capabilities including zero-shot learning, high fidelity data synthesis, and out of domain generalization. However, the parameter capacity of FMs is still limited, leading to poor out-of-the-box performance of FMs on many expert tasks (e.g. retrieval of car manuals technical illustrations from language queries), data for which is either unseen or belonging to a long-tail part of the data distribution of the huge datasets used for FM pre-training. This underlines the necessity to explicitly evaluate and finetune FMs on such expert tasks, arguably ones that appear the most in practical real-world applications. In this paper, we propose a first of its kind FETA benchmark built around the task of teaching FMs to understand technical documentation, via learning to match their graphical illustrations to corresponding language descriptions. Our FETA benchmark focuses on text-to-image and image-to-text retrieval in public car manuals and sales catalogue brochures. FETA is equipped with a procedure for completely automatic annotation extraction (code would be released upon acceptance), allowing easy extension of FETA to more documentation types and application domains in the future. Our automatic annotation leads to an automated performance metric shown to be consistent with metrics computed on human-curated annotations (also released). We provide multiple baselines and analysis of popular FMs on FETA leading to several interesting findings that we believe would be very valuable to the FM community, paving the way towards real-world application of FMs for many practical expert tasks currently being `overlooked' by standard benchmarks focusing on common objects.

## [pFL-Bench: A Comprehensive Benchmark for Personalized Federated Learning](#)

- Daoyuan Chen · Dawei Gao · Weirui Kuang · Yaliang Li · Bolin Ding
- abstract@[open-review](#): Personalized Federated Learning (pFL), which utilizes and deploys distinct local models, has gained increasing attention in recent years due to its success in handling the statistical heterogeneity of FL clients. However, standardized evaluation and systematical analysis of diverse pFL methods remain a challenge. Firstly, the highly varied datasets, FL simulation settings and pFL implementations prevent easy and fair comparisons of pFL methods. Secondly, the current pFL literature diverges in the adopted evaluation and ablation protocols. Finally, the effectiveness and robustness of pFL methods are under-explored in various practical scenarios, such as the generalization to new clients and the participation of resource-limited clients. To tackle these challenges, we propose the first comprehensive pFL benchmark, pFL-Bench, for facilitating rapid, reproducible, standardized and thorough pFL evaluation. The proposed benchmark contains more than 10 dataset variants in various application domains with a unified data partition and realistic heterogeneous settings; a modular and easy-to-extend pFL codebase with more than 20 competitive pFL method implementations; and systematic evaluations under containerized environments in terms of generalization, fairness, system overhead, and convergence. We highlight the benefits and potential of state-of-the-art pFL methods and hope pFL-Bench enables further pFL research and broad applications that would otherwise be difficult owing to the absence of a dedicated benchmark. The code is released at <https://github.com/alibaba/FederatedScope/tree/master/benchmark/pFL-Bench>.

## [OccGen: Selection of Real-world Multilingual Parallel Data Balanced in Gender within Occupations](#)

- Marta Costa-jussÃ· · Christine Basta · Oriol Domingo · AndrÃ© Rubungo
- abstract@[open-review](#): This paper describes the OCCGEN toolkit, which allows extracting multilingual parallel data balanced in gender within occupations. OCCGEN can extract datasets that reflect gender diversity (beyond binary) more fairly in society to be further used to explicitly mitigate occupational gender stereotypes. We propose two use cases that extract evaluation datasets for machine translation in four high-resource languages from different linguistic families and in a low-resource African language. Our analysis of these use cases shows that translation outputs in high-resource languages tend to worsen in feminine subsets (compared to masculine). This can be explained because less attention is paid to the source sentence. Then, more attention is given to the target prefix overgeneralizing to the most frequent masculine forms.

## [A Dataset for Efforts Towards Achieving the Sustainable Development Goal of Safe Working Environments](#)

- Eirik Lund Flogard · Ole Jakob Mengshoel
- abstract@[open-review](#): Among United Nations' 17 Sustainable Development Goals (SDGs), we highlight SDG 8 on Decent Work and Economic Growth. Specifically, we consider how to achieve subgoal 8.8, "protect labour rights and promote safe working environments for all workers [...]", in light of poor health, safety and environment (HSE) conditions being a widespread problem at workplaces. In EU alone, it is estimated that more than 4000 deaths occur each year due to poor working conditions. To handle the problem and achieve SDG 8, governmental agencies conduct labour inspections and it is therefore essential that these are carried out efficiently. Current research suggests that machine learning (ML) can be used to improve labour inspections,

for instance by selecting organisations for inspections more effectively. However, the research in this area is very limited, in part due to a lack of publicly available data. Consequently, we introduce a new dataset called the Labour Inspection Checklists Dataset (LICD), which we have made publicly available. LICD consists of 63634 instances where each instance is an inspection conducted by the Norwegian Labour Inspection Authority. LICD has 575 features and two potential target variables: checklists and non-compliance. The dataset provides several ML research opportunities; we discuss two demonstration experiments. One experiment deals with the problem of selecting a relevant checklist for inspecting a given target organisation. The other experiment concerns the problem of predicting HSE violations, given a specific checklist and a target organisation. Our experimental results, while promising, suggest that achieving good ML classification performance is difficult for both problems. This motivates future research to improve ML performance, inspire other data analysis efforts, and ultimately achieve SDG 8.

## [EgoTaskQA: Understanding Human Tasks in Egocentric Videos](#)

- Baoxiong Jia · Ting Lei · Song-Chun Zhu · Siyuan Huang
- abstract@[open-review](#): Understanding human tasks through video observations is an essential capability of intelligent agents. The challenges of such capability lie in the difficulty of generating a detailed understanding of situated actions, their effects on object states (ie, state changes), and their causal dependencies. These challenges are further aggravated by the natural parallelism from multi-tasking and partial observations in multi-agent collaboration. Most prior works leverage action localization or future prediction as an \textit{indirect} metric for evaluating such task understanding from videos. To make a \textit{direct} evaluation, we introduce the EgoTaskQA benchmark that provides a single home for the crucial dimensions of task understanding through question answering on real-world egocentric videos. We meticulously design questions that target the understanding of (1) action dependencies and effects, (2) intents and goals, and (3) agents' beliefs about others. These questions are divided into four types, including descriptive (what status?), predictive (what will?), explanatory (what caused?), and counterfactual (what if?) to provide diagnostic analyses on \textit{spatial, temporal, and causal} understandings of goal-oriented tasks. We evaluate state-of-the-art video reasoning models on our benchmark and show their significant gaps between humans in understanding complex goal-oriented egocentric videos. We hope this effort would drive the vision community to move onward with goal-oriented video understanding and reasoning.

## [Honor of Kings Arena: an Environment for Generalization in Competitive Reinforcement Learning](#)

- Hua Wei · Jingxiao Chen · Xiyang Ji · Hongyang Qin · Minwen Deng · Siqin Li · Liang Wang · Weinan Zhang · Yong Yu · Liu Linc · Lanxiao Huang · Deheng Ye · Qiang Fu · Wei Yang
- abstract@[open-review](#): This paper introduces Honor of Kings Arena, a reinforcement learning (RL) environment based on the Honor of Kings, one of the world's most popular games at present. Compared to other environments studied in most previous work, ours presents new generalization challenges for competitive reinforcement learning. It is a multi-agent problem with one agent competing against its opponent; and it requires the generalization ability as it has diverse targets to control and diverse opponents to compete with. We describe the observation, action, and reward specifications for the Honor of Kings domain and provide an open-source Python-based interface for communicating with the game engine. We provide twenty target heroes with a variety of tasks in Honor of Kings Arena and present initial baseline results for RL-based methods with feasible computing resources. Finally, we showcase the generalization challenges imposed by Honor of Kings Arena and possible remedies to the challenges. All of the software, including the environment-class, are publicly available.

## [MTNeuro: A Benchmark for Evaluating Representations of Brain Structure Across Multiple Levels of Abstraction](#)

- Jorge Quesada · Lakshmi Sathidevi · Ran Liu · Nauman Ahad · Joy Jackson · Mehdi Azabou · Jingyun Xiao · Christopher Liding · Matthew Jin · Carolina Urzay · William Gray-Roncal · Erik Johnson · Eva Dyer
- abstract@[open-review](#): There are multiple scales of abstraction from which we can describe the same image, depending on whether we are focusing on fine-grained details or a more global attribute of the image. In brain mapping, learning to automatically parse images to build representations of both small-scale features (e.g., the presence of cells or blood vessels) and global properties of an image (e.g., source brain region) is a crucial and open challenge. However, most existing datasets and benchmarks for neuroanatomy consider only a single downstream task at a time. We introduce a new dataset, annotations, and multiple downstream tasks that provide diverse ways to readout information about brain structure and architecture from the same image. Our multi-task neuroimaging benchmark (MTNeuro) is built on volumetric, micrometer-resolution X-ray microtomography imaging of a large thalamocortical section of mouse brain, encompassing multiple cortical and subcortical regions, that reveals dense reconstructions of the underlying microstructure (i.e., cell bodies, vasculature, and axons). We generated a number of different prediction challenges and evaluated several supervised and self-supervised models for brain-region prediction and pixel-level semantic segmentation of microstructures. Our experiments not only highlight the rich heterogeneity of this dataset, but also provide insights into how self-supervised approaches can be used to learn representations that capture multiple attributes of a single image and perform well on a variety of downstream tasks. Datasets, code, and pre-trained baseline models are provided at: <https://mtneuro.github.io/>.

## [SafeBench: A Benchmarking Platform for Safety Evaluation of Autonomous Vehicles](#)

- Chejian Xu · Wenhao Ding · Weijie Lyu · ZUXIN LIU · Shuai Wang · Yihan He · Hanjiang Hu · DING ZHAO · Bo Li
- abstract@[open-review](#): As shown by recent studies, machine intelligence-enabled systems are vulnerable to test cases resulting from either adversarial manipulation or natural distribution shifts. This has raised great concerns about deploying machine learning algorithms for real-world applications, especially in safety-critical domains such as autonomous driving (AD). On the other hand, traditional AD testing on naturalistic scenarios requires hundreds of millions of driving miles due to the high dimensionality and rareness of the safety-critical scenarios in the real world. As a result, several approaches for autonomous driving evaluation have been explored, which are usually, however, based on different simulation platforms, types of safety-critical scenarios, scenario generation algorithms, and driving route variations. Thus, despite a large amount of effort in autonomous driving testing, it is still challenging to compare and understand the effectiveness and efficiency of different testing scenario generation algorithms and testing mechanisms under similar conditions. In this paper, we aim to provide the first unified platform SafeBench to integrate different types of safety-critical testing scenarios, scenario generation algorithms, and other variations such as driving routes and environments. In particular, we consider 8 safety-critical testing scenarios following National Highway Traffic Safety Administration (NHTSA) and develop 4 scenario generation algorithms considering 10 variations for each scenario. Meanwhile, we implement 4 deep reinforcement learning-based AD algorithms with 4 types of input (e.g., bird's-eye view, camera) to perform fair comparisons on SafeBench. We find our generated testing scenarios are indeed more challenging and observe the trade-off between the performance of AD agents under benign and safety-critical testing scenarios. We believe our unified platform SafeBench for large-scale and effective autonomous driving testing will motivate the development of new testing scenario generation and safe AD algorithms. SafeBench is available at <https://safebench.github.io/>.

## [AnoShift: A Distribution Shift Benchmark for Unsupervised Anomaly Detection](#)

- Marius Dragoi · Elena Burceanu · Emanuela Haller · Andrei Manolache · Florin Brad
- abstract@[open-review](#): Analyzing the distribution shift of data is a growing research direction in nowadays Machine Learning (ML), leading to emerging new benchmarks that focus on providing a suitable scenario for studying the generalization properties of ML models. The existing benchmarks are focused on supervised learning, and to the best of our knowledge, there is none for unsupervised learning. Therefore, we introduce an unsupervised anomaly detection benchmark with data that shifts over time, built over Kyoto-2006+, a traffic dataset for network intrusion detection. This type of data meets the premise of shifting the input distribution: it covers a large time span (10 years), with naturally occurring changes over time (e.g. users modifying their behavior patterns, and software updates). We first highlight the non-stationary nature of the data, using a basic per-feature analysis, t-SNE, and an

Optimal Transport approach for measuring the overall distribution distances between years. Next, we propose AnoShift, a protocol splitting the data in IID, NEAR, and FAR testing splits. We validate the performance degradation over time with diverse models, ranging from classical approaches to deep learning. Finally, we show that by acknowledging the distribution shift problem and properly addressing it, the performance can be improved compared to the classical training which assumes independent and identically distributed data (on average, by up to 3% for our approach). Dataset and code are available at <https://github.com/bit-ml/AnoShift/>.

## [Why do tree-based models still outperform deep learning on typical tabular data?](#)

- Leo Grinsztajn · Edouard Oyallon · Gael Varoquaux
- abstract@[open-review](#): While deep learning has enabled tremendous progress on text and image datasets, its superiority on tabular data is not clear. We contribute extensive benchmarks of standard and novel deep learning methods as well as tree-based models such as XGBoost and Random Forests, across a large number of datasets and hyperparameter combinations. We define a standard set of 45 datasets from varied domains with clear characteristics of tabular data and a benchmarking methodology accounting for both fitting models and finding good hyperparameters. Results show that tree-based models remain state-of-the-art on medium-sized data ( $\sim 10K$  samples) even without accounting for their superior speed. To understand this gap, we conduct an empirical investigation into the differing inductive biases of tree-based models and neural networks. This leads to a series of challenges which should guide researchers aiming to build tabular-specific neural network: 1) be robust to uninformative features, 2) preserve the orientation of the data, and 3) be able to easily learn irregular functions. To stimulate research on tabular architectures, we contribute a standard benchmark and raw data for baselines: every point of a 20,000 compute hours hyperparameter search for each learner.

## [Addressing Resource Scarcity across Sign Languages with Multilingual Pretraining and Unified-Vocabulary Datasets](#)

- Gokul NC · Manideep Ladi · Sumit Negi · Prem Selvaraj · Pratyush Kumar · Mitesh Khapra
- abstract@[open-review](#): There are over 300 sign languages in the world, many of which have very limited or no labelled sign-to-text datasets. To address low-resource data scenarios, self-supervised pretraining and multilingual finetuning have been shown to be effective in natural language and speech processing. In this work, we apply these ideas to sign language recognition. We make three contributions.- First, we release SignCorpus, a large pretraining dataset on sign languages comprising about 4.6K hours of signing data across 10 sign languages. SignCorpus is curated from sign language videos on the internet, filtered for data quality, and converted into sequences of pose keypoints thereby removing all personal identifiable information (PII).- Second, we release Sign2Vec, a graph-based model with 5.2M parameters that is pretrained on SignCorpus. We envisage Sign2Vec as a multilingual large-scale pretrained model which can be fine-tuned for various sign recognition tasks across languages.- Third, we create MultiSign-ISLR -- a multilingual and label-aligned dataset of sequences of pose keypoints from 11 labelled datasets across 7 sign languages, and MultiSign-FS -- a new finger-spelling training and test set across 7 languages. On these datasets, we fine-tune Sign2Vec to create multilingual isolated sign recognition models. With experiments on multiple benchmarks, we show that pretraining and multilingual transfer are effective giving significant gains over state-of-the-art results. All datasets, models, and code has been made open-source via the OpenHands toolkit.

## [OpenXAI: Towards a Transparent Evaluation of Model Explanations](#)

- Chirag Agarwal · Satyapriya Krishna · Eshika Saxena · Martin Pawelczyk · Nari Johnson · Isha Puri · Marinka Zitnik · Himabindu Lakkaraju
- abstract@[open-review](#): While several types of post hoc explanation methods have been proposed in recent literature, there is very little work on systematically benchmarking these methods. Here, we introduce OpenXAI, a comprehensive and extensible open-source framework for evaluating and benchmarking post hoc explanation methods. OpenXAI comprises of the following key components: (i) a flexible synthetic data generator and a collection of diverse real-world datasets, pre-trained models, and state-of-the-art feature attribution methods, (ii) open-source implementations of twenty-two quantitative metrics for evaluating faithfulness, stability (robustness), and fairness of explanation methods, and (iii) the first ever public XAI leaderboards to readily compare several explanation methods across a wide variety of metrics, models, and datasets. OpenXAI is easily extensible, as users can readily evaluate custom explanation methods and incorporate them into our leaderboards. Overall, OpenXAI provides an automated end-to-end pipeline that not only simplifies and standardizes the evaluation of post hoc explanation methods, but also promotes transparency and reproducibility in benchmarking these methods. While the first release of OpenXAI supports only tabular datasets, the explanation methods and metrics that we consider are general enough to be applicable to other data modalities. OpenXAI datasets and data loaders, implementations of state-of-the-art explanation methods and evaluation metrics, as well as leaderboards are publicly available at <https://open-xai.github.io/>. OpenXAI will be regularly updated to incorporate text and image datasets, other new metrics and explanation methods, and welcomes inputs from the community.

## [Turning the Tables: Biased, Dynamic, Imbalanced Tabular Datasets for ML Research](#)

- Sárgio Jesus · José Pombal · Duarte Alves · André Cruz · Pedro Saleiro · Rita Ribeiro · João Gama · Pedro Bizarro
- abstract@[open-review](#): Evaluating new techniques on realistic datasets plays a crucial role in the development of ML research and its broader adoption by practitioners. In recent years, there has been a significant increase of publicly available unstructured data resources for computer vision and NLP tasks. However, tabular data – which is prevalent in many high-stakes decision-making domains – has been lagging behind. To bridge this gap, we present Bank Account Fraud (BAF), the first publicly available privacy-preserving, large-scale, realistic suite of tabular datasets. The suite was generated by applying state-of-the-art tabular data generation techniques on an anonymized, real-world bank account opening fraud detection dataset. This setting carries a set of challenges that are commonplace in real-world applications, including temporal dynamics and significant class imbalance. Additionally, to allow practitioners to stress test both performance and fairness of ML methods, each dataset variant of BAF features specific types of data bias, including time-related patterns. With this resource, we aim to provide the ML and Fair ML research communities with a more realistic, complete, and robust test bed to evaluate novel and existing methods.

## [mRI: Multi-modal 3D Human Pose Estimation Dataset using mmWave, RGB-D, and Inertial Sensors](#)

- Sizhe An · Yin Li · Umit Ogras
- abstract@[open-review](#): The ability to estimate 3D human body pose and movement, also known as human pose estimation (HPE), enables many applications for home-based health monitoring, such as remote rehabilitation training. Several possible solutions have emerged using sensors ranging from RGB cameras, depth sensors, millimeter-Wave (mmWave) radars, and wearable inertial sensors. Despite previous efforts on datasets and benchmarks for HPE, few dataset exploits multiple modalities and focuses on home-based health monitoring. To bridge the gap, we present mRI, a multi-modal 3D human pose estimation dataset with mmWave, RGB-D, and Inertial Sensors. Our dataset consists of over 160k synchronized frames from 20 subjects performing rehabilitation exercises and supports the benchmarks of HPE and action detection. We perform extensive experiments using our dataset and delineate the strength of each modality. We hope that the release of mRI can catalyze the research in pose estimation, multi-modal learning, and action understanding, and more importantly facilitate the applications of home-based health monitoring.

## [TweetNERD - End to End Entity Linking Benchmark for Tweets](#)

- Shubhanshu Mishra · Aman Saini · Raheleh Makki · Sneha Mehta · Aria Haghghi · Ali Mollahosseini
- abstract@[open-review](#): Named Entity Recognition and Disambiguation (NERD) systems are foundational for information retrieval, question answering, event detection, and other natural language processing (NLP) applications. We introduce TweetNERD, a dataset of 340K+ Tweets across 2010-2021, for benchmarking NERD systems on Tweets. This is the largest and most temporally diverse open sourced dataset benchmark for NERD on Tweets and can be used to facilitate research in this area. We describe evaluation setup with TweetNERD for three NERD tasks: Named Entity Recognition (NER), Entity

Linking with True Spans (EL), and End to End Entity Linking (End2End); and provide performance of existing publicly available methods on specific TweetNERD splits. TweetNERD is available at: <https://doi.org/10.5281/zenodo.6617192> under Creative Commons Attribution 4.0 International (CC BY 4.0) license. Check out more details at <https://github.com/twitter-research/TweetNERD>.

## [Change Event Dataset for Discovery from Spatio-temporal Remote Sensing Imagery](#)

- Utkarsh Mall · Bharath Hariharan · Kavita Bala
- abstract@[open-review](#): Satellite imagery is increasingly available, high resolution, and temporally detailed. Changes in spatio-temporal datasets such as satellite images are particularly interesting as they reveal the many events and forces that shape our world. However, finding such interesting and meaningful change events from the vast data is challenging. In this paper, we present new datasets for such change events that include semantically meaningful events like road construction. Instead of manually annotating the very large corpus of satellite images, we introduce a novel unsupervised approach that takes a large spatio-temporal dataset from satellite images and finds interesting change events. To evaluate the meaningfulness on these datasets we create 2 benchmarks namely CaiRoad and CalFire which capture the events of road construction and forest fires. These new benchmarks can be used to evaluate semantic retrieval/classification performance. We explore these benchmarks qualitatively and quantitatively by using several methods and show that these new datasets are indeed challenging for many existing methods.

## [PROSPECT: Labeled Tandem Mass Spectrometry Dataset for Machine Learning in Proteomics](#)

- Omar Shouman · Wassim Gabriel · Victor-George Giurcoiu · Vitor Sternlicht · Mathias Wilhelm
- abstract@[open-review](#): Proteomics is the interdisciplinary field focusing on the large-scale study of proteins. Proteins essentially organize and execute all functions within organisms. Today, the bottom-up analysis approach is the most commonly used workflow, where proteins are digested into peptides and subsequently analyzed using Tandem Mass Spectrometry (MS/MS). MS-based proteomics has transformed various fields in life sciences, such as drug discovery and biomarker identification. Today, proteomics is entering a phase where it is helpful for clinical decision-making. Computational methods are vital in turning large amounts of acquired raw MS data into information and, ultimately, knowledge. Deep learning has proved its success in multiple domains as a robust framework for supervised and unsupervised machine learning problems. In proteomics, scientists are increasingly leveraging the potential of deep learning to predict the properties of peptides based on their sequence to improve their confident identification. However, a reference dataset is missing, covering several proteomics tasks, enabling performance comparison, and evaluating reproducibility and generalization. Here, we present a large labeled proteomics dataset spanning several tasks in the domain to address this challenge. We focus on two common applications: peptide retention time and MS/MS spectrum prediction. We review existing methods and task formulations from a machine learning perspective and recommend suitable evaluation metrics and visualizations. With an accessible dataset, we aim to lower the entry barrier and enable faster development in machine learning for proteomics.

## [Tenrec: A Large-scale Multipurpose Benchmark Dataset for Recommender Systems](#)

- guanghu yuan · Fajie Yuan · Yudong Li · Beibei Kong · Shujie Li · Lei Chen · Min Yang · Chenyun YU · Bo Hu · Zang Li · Yu Xu · Xiaohu Qie
- abstract@[open-review](#): Existing benchmark datasets for recommender systems (RS) either are created at a small scale or involve very limited forms of user feedback. RS models evaluated on such datasets often lack practical values for large-scale real-world applications. In this paper, we describe Tenrec, a novel and publicly available data collection for RS that records various user feedback from four different recommendation scenarios. To be specific, Tenrec has the following five characteristics: (1) it is large-scale, containing around 5 million users and 140 million interactions; (2) it has not only positive user feedback, but also true negative feedback (vs. one-class recommendation); (3) it contains overlapped users and items across four different scenarios; (4) it contains various types of user positive feedback, in forms of clicking, liking, sharing, and following, etc; (5) it contains additional features beyond the user IDs and item IDs. We verify Tenrec on ten diverse recommendation tasks by running several classical baseline models per task. Tenrec has the potential to become a useful benchmark dataset for a majority of popular recommendation tasks. Our source codes and datasets will be included in supplementary materials.

## [ConfLab: A Data Collection Concept, Dataset, and Benchmark for Machine Analysis of Free-Standing Social Interactions in the Wild](#)

- Chirag A Raman · Jose Vargas Quiros · Stephanie Tan · Ashraful Islam · Ekin Gedik · Hayley Hung
- abstract@[open-review](#): Recording the dynamics of unscripted human interactions in the wild is challenging due to the delicate trade-offs between several factors: participant privacy, ecological validity, data fidelity, and logistical overheads. To address these, following a 'datasets for the community by the community' ethos, we propose the Conference Living Lab (ConfLab): a new concept for multimodal multisensor data collection of in-the-wild free-standing social conversations. For the first instantiation of ConfLab described here, we organized a real-life professional networking event at a major international conference. Involving 48 conference attendees, the dataset captures a diverse mix of status, acquaintance, and networking motivations. Our capture setup improves upon the data fidelity of prior in-the-wild datasets while retaining privacy sensitivity: 8 videos (1920x1080, 60 fps) from a non-invasive overhead view, and custom wearable sensors with onboard recording of body motion (full 9-axis IMU), privacy-preserving low-frequency audio (1250 Hz), and Bluetooth-based proximity. Additionally, we developed custom solutions for distributed hardware synchronization at acquisition, and time-efficient continuous annotation of body keypoints and actions at high sampling rates. Our benchmarks showcase some of the open research tasks related to in-the-wild privacy-preserving social data analysis: keypoints detection from overhead camera views, skeleton-based no-audio speaker detection, and F-formation detection.

## [DC-BENCH: Dataset Condensation Benchmark](#)

- Justin CUI · Ruochen Wang · Si Si · Cho-Jui Hsieh
- abstract@[open-review](#): Dataset Condensation is a newly emerging technique aiming at learning a tiny dataset that captures the rich information encoded in the original dataset. As the size of datasets contemporary machine learning models rely on becomes increasingly large, condensation methods become a prominent direction for accelerating network training and reducing data storage. Despite numerous methods have been proposed in this rapidly growing field, evaluating and comparing different condensation methods is non-trivial and still remains an open issue. The quality of condensed dataset are often shadowed by many critical contributing factors to the end performance, such as data augmentation and model architectures. The lack of a systematic way to evaluate and compare condensation methods not only hinders our understanding of existing techniques, but also discourages practical usage of the synthesized datasets. This work provides the first large-scale standardized benchmark on Dataset Condensation. It consists of a suite of evaluations to comprehensively reflect the generability and effectiveness of condensation methods through the lens of their generated dataset. Leveraging this benchmark, we conduct a large-scale study of current condensation methods, and report many insightful findings that open up new possibilities for future development. The benchmark library, including evaluators, baseline methods, and generated datasets, is open-sourced to facilitate future research and application.

## [DABS 2.0: Improved Datasets and Algorithms for Universal Self-Supervision](#)

- Alex Tamkin · Gaurab Banerjee · Mohamed Owda · Vincent Liu · Shashank Rammoorthy · Noah Goodman
- abstract@[open-review](#): Universal self-supervised (SSL) algorithms hold enormous promise for making machine learning accessible to high-impact domains such as protein biology, manufacturing, and genomics. We present DABS 2.0: a set of improved datasets and algorithms for advancing research

on universal SSL. We extend the recently-introduced DABS benchmark with the addition of five real-world science and engineering domains: protein biology, bacterial genomics, multispectral satellite imagery, semiconductor wafers, and particle physics, bringing the total number of domains in the benchmark to twelve. We also propose a new universal SSL algorithm, Capri, and a generalized version of masked autoencoding, and apply both on all twelve domains---the most wide-ranging exploration of SSL yet. We find that multiple algorithms show gains across domains, outperforming previous baselines. In addition, we demonstrate the usefulness of DABS for scientific study of SSL by investigating the optimal corruption rate for each algorithm, showing that the best setting varies based on the domain. Code will be released at <http://github.com/alextamkin/dabs>

## [Communicating Natural Programs to Humans and Machines](#)

- Sam Acquaviva · Yewen Pu · Marta Kryven · Theodoros Sechopoulos · Catherine Wong · Gabrielle Ecanow · Maxwell Nye · Michael Tessler · Josh Tenenbaum
- abstract@[open-review](#): The Abstraction and Reasoning Corpus (ARC) is a set of procedural tasks that tests an agent's ability to flexibly solve novel problems. While most ARC tasks are easy for humans, they are challenging for state-of-the-art AI. What makes building intelligent systems that can generalize to novel situations such as ARC difficult? We posit that the answer might be found by studying the difference of \$\textit{language}\$: While humans readily generate and interpret instructions in a general language, computer systems are shackled to a narrow domain-specific language that they can precisely execute. We present LARC, the \$\textit{Language-complete ARC}\$: a collection of natural language descriptions by a group of human participants who instruct each other on how to solve ARC tasks using language alone, which contains successful instructions for 88% of the ARC tasks. We analyze the collected instructions as 'natural programs', finding that while they resemble computer programs, they are distinct in two ways: First, they contain a wide range of primitives; Second, they frequently leverage communicative strategies beyond directly executable codes. We demonstrate that these two distinctions prevent current program synthesis techniques from leveraging LARC to its full potential, and give concrete suggestions on how to build the next-generation program synthesizers.

## [Hard ImageNet: Segmentations for Objects with Strong Spurious Cues](#)

- Mazda Moayeri · Sahil Singla · Soheil Feizi
- abstract@[open-review](#): Deep classifiers are known to rely on spurious features, leading to reduced generalization. The severity of this problem varies significantly by class. We identify 15 classes in ImageNet with very strong spurious cues, and collect segmentation masks for these challenging objects to form 'Hard ImageNet'. Leveraging noise, saliency, and ablation based metrics, we demonstrate that models rely on spurious features in Hard ImageNet far more than in RIVAL10, an ImageNet analog to CIFAR10. We observe Hard ImageNet objects are less centered and occupy much less space in their images than RIVAL10 objects, leading to greater spurious feature reliance. Further, we use robust neural features to automatically rank our images based on the degree of spurious cues present. Comparing images with high and low rankings within a class reveals the exact spurious features models rely upon, and shows reduced performance when spurious features are absent. With Hard ImageNet's image rankings, object segmentations, and our extensive evaluation suite, the community can begin to address the problem of learning to detect challenging objects 'for the right reasons', despite the presence of strong spurious cues.

## [The Surprising Effectiveness of PPO in Cooperative Multi-Agent Games](#)

- Chao Yu · Akash Velu · Eugene Vinitsky · Jiaxuan Gao · Yu Wang · Alexandre Bayen · YI WU
- abstract@[open-review](#): Proximal Policy Optimization (PPO) is a ubiquitous on-policy reinforcement learning algorithm but is significantly less utilized than off-policy learning algorithms in multi-agent settings. This is often due to the belief that PPO is significantly less sample efficient than off-policy methods in multi-agent systems. In this work, we carefully study the performance of PPO in cooperative multi-agent settings. We show that PPO-based multi-agent algorithms achieve surprisingly strong performance in four popular multi-agent testbeds: the particle-world environments, the StarCraft multi-agent challenge, the Hanabi challenge, and Google Research Football, with minimal hyperparameter tuning and without any domain-specific algorithmic modifications or architectures. Importantly, compared to competitive off-policy methods, PPO often achieves competitive or superior results in both final returns and sample efficiency. Finally, through ablation studies, we analyze implementation and hyperparameter factors that are critical to PPO's empirical performance, and give concrete practical suggestions regarding these factors. Our results show that when using these practices, simple PPO-based methods are a strong baseline in cooperative multi-agent reinforcement learning. Source code is released at <https://github.com/marlbenchmark/on-policy>.

## [CAESAR: An Embodied Simulator for Generating Multimodal Referring Expression Datasets](#)

- Md Mofijul Islam · Reza Mirzaee · Alexi Gladstone · Haley Green · Tariq Iqbal
- abstract@[open-review](#): Humans naturally use verbal utterances and nonverbal gestures to refer to various objects (known as '\$\textit{referring expressions}\$') in different interactional scenarios. As collecting real human interaction datasets are costly and laborious, synthetic datasets are often used to train models to unambiguously detect relationships among objects. However, existing synthetic data generation tools that provide referring expressions generally neglect nonverbal gestures. Additionally, while a few small-scale datasets contain multimodal cues (verbal and nonverbal), these datasets only capture the nonverbal gestures from an exo-centric perspective (observer). As models can use complementary information from multimodal cues to recognize referring expressions, generating multimodal data from multiple views can help to develop robust models. To address these critical issues, in this paper, we present a novel embodied simulator, CAESAR, to generate multimodal referring expressions containing both verbal utterances and nonverbal cues captured from multiple views. Using our simulator, we have generated two large-scale embodied referring expression datasets, which we will release publicly. We have conducted experimental analyses on embodied spatial relation grounding using various state-of-the-art baseline models. Our experimental results suggest that visual perspective affects the models' performance; and that nonverbal cues improve spatial relation grounding accuracy. Finally, we will release the simulator publicly to allow researchers to generate new embodied interaction datasets.

## [Long Range Graph Benchmark](#)

- Vijay Prakash Dwivedi · Ladislav Rampářek · Mikhail Galkin · Ali Parviz · Guy Wolf · Anh Tuan Luu · Dominique Beaini
- abstract@[open-review](#): Graph Neural Networks (GNNs) that are based on the message passing (MP) paradigm generally exchange information between 1-hop neighbors to build node representations at each layer. In principle, such networks are not able to capture long-range interactions (LRI) that may be desired or necessary for learning a given task on graphs. Recently, there has been an increasing interest in development of Transformer-based methods for graphs that can consider full node connectivity beyond the original sparse structure, thus enabling the modeling of LRI. However, MP-GNNs that simply rely on 1-hop message passing often fare better in several existing graph benchmarks when combined with positional feature representations, among other innovations, hence limiting the perceived utility and ranking of Transformer-like architectures. Here, we present the Long Range Graph Benchmark (LRGB) with 5 graph learning datasets: '\$\textit{PascalVOC-SP}\$', '\$\textit{COCO-SP}\$', '\$\textit{PCQM-Contact}\$', '\$\textit{Peptides-func}\$' and '\$\textit{Peptides-struct}\$' that arguably require LRI reasoning to achieve strong performance in a given task. We benchmark both baseline GNNs and Graph Transformer networks to verify that the models which capture long-range dependencies perform significantly better on these tasks. Therefore, these datasets are suitable for benchmarking and exploration of MP GNNs and Graph Transformer architectures that are intended to capture LRI.

## [SoundSpaces 2.0: A Simulation Platform for Visual-Acoustic Learning](#)

- Changan Chen · Carl Schissler · Sanchit Garg · Philip Kobernik · Alexander Clegg · Paul Calamia · Dhruv Batra · Philip Robinson · Kristen Grauman

- abstract@[open-review](#): We introduce SoundSpaces 2.0, a platform for on-the-fly geometry-based audio rendering for 3D environments. Given a 3D mesh of a real-world environment, SoundSpaces can generate highly realistic acoustics for arbitrary sounds captured from arbitrary microphone locations. Together with existing 3D visual assets, it supports an array of audio-visual research tasks, such as audio-visual navigation, mapping, source localization and separation, and acoustic matching. Compared to existing resources, SoundSpaces 2.0 has the advantages of allowing continuous spatial sampling, generalization to novel environments, and configurable microphone and material properties. To our knowledge, this is the first geometry-based acoustic simulation that offers high fidelity and realism while also being fast enough to use for embodied learning. We showcase the simulator's properties and benchmark its performance against real-world audio measurements. In addition, we demonstrate two downstream tasks---embodied navigation and far-field automatic speech recognition---and highlight sim2real performance for the latter. SoundSpaces 2.0 is publicly available to facilitate wider research for perceptual systems that can both see and hear.

## [GLOBEM Dataset: Multi-Year Datasets for Longitudinal Human Behavior Modeling Generalization](#)

- Xuhai Xu · Han Zhang · Yasaman Sefidgar · Yiyi Ren · Xin Liu · Woosuk Seo · Jennifer Brown · Kevin Kuehn · Mike Merrill · Paula Nurius · Shwetak Patel · Tim Althoff · Margaret Morris · Eve Riskin · Jennifer Mankoff · Anind Dey
- abstract@[open-review](#): Recent research has demonstrated the capability of behavior signals captured by smartphones and wearables for longitudinal behavior modeling. However, there is a lack of a comprehensive public dataset that serves as an open testbed for fair comparison among algorithms. Moreover, prior studies mainly evaluate algorithms using data from a single population within a short period, without measuring the cross-dataset generalizability of these algorithms. We present the first multi-year passive sensing datasets, containing over 700 user-years and 497 unique users™ data collected from mobile and wearable sensors, together with a wide range of well-being metrics. Our datasets can support multiple cross-dataset evaluations of behavior modeling algorithms™ generalizability across different users and years. As a starting point, we provide the benchmark results of 18 algorithms on the task of depression detection. Our results indicate that both prior depression detection algorithms and domain generalization techniques show potential but need further research to achieve adequate cross-dataset generalizability. We envision our multi-year datasets can support the ML community in developing generalizable longitudinal behavior modeling algorithms.

## [MineDojo: Building Open-Ended Embodied Agents with Internet-Scale Knowledge](#)

- Linxi Fan · Guanzhi Wang · Yunfan Jiang · Ajay Mandlekar · Yuncong Yang · Haoyi Zhu · Andrew Tang · De-An Huang · Yuke Zhu · Anima Anandkumar
- abstract@[open-review](#): Autonomous agents have made great strides in specialist domains like Atari games and Go. However, they typically learn tabula rasa in isolated environments with limited and manually conceived objectives, thus failing to generalize across a wide spectrum of tasks and capabilities. Inspired by how humans continually learn and adapt in the open world, we advocate a trinity of ingredients for building generalist agents: 1) an environment that supports a multitude of tasks and goals, 2) a large-scale database of multimodal knowledge, and 3) a flexible and scalable agent architecture. We introduce MineDojo, a new framework built on the popular Minecraft game that features a simulation suite with thousands of diverse open-ended tasks and an internet-scale knowledge base with Minecraft videos, tutorials, wiki pages, and forum discussions. Using MineDojo's data, we propose a novel agent learning algorithm that leverages large pre-trained video-language models as a learned reward function. Our agent is able to solve a variety of open-ended tasks specified in free-form language without any manually designed dense shaping reward. We open-source the simulation suite, knowledge bases, algorithm implementation, and pretrained models (<https://minedojo.org>) to promote research towards the goal of generally capable embodied agents.

## [ComMU: Dataset for Combinatorial Music Generation](#)

- Lee Hyun · Taehyun Kim · Hyolim Kang · Minjoo Ki · Hyeonchan Hwang · kwanho park · Sharang Han · Seon Joo Kim
- abstract@[open-review](#): Commercial adoption of automatic music composition requires the capability of generating diverse and high-quality music suitable for the desired context (e.g., music for romantic movies, action games, restaurants, etc.). In this paper, we introduce combinatorial music generation, a new task to create varying background music based on given conditions. Combinatorial music generation creates short samples of music with rich musical metadata, and combines them to produce a complete music. In addition, we introduce ComMU, the first symbolic music dataset consisting of short music samples and their corresponding 12 musical metadata for combinatorial music generation. Notable properties of ComMU are that (1) dataset is manually constructed by professional composers with an objective guideline that induces regularity, and (2) it has 12 musical metadata that embraces composers' intentions. Our results show that we can generate diverse high-quality music only with metadata, and that our unique metadata such as track-role and extended chord quality improves the capacity of the automatic composition. We highly recommend watching our video before reading the paper (<https://pozalabs.github.io/ComMU/>).

## [SCAMPS: Synthetics for Camera Measurement of Physiological Signals](#)

- Daniel McDuff · Miah Wander · Xin Liu · Brian Hill · Javier Hernandez · Jonathan Lester · Tadas Baltrusaitis
- abstract@[open-review](#): The use of cameras and computational algorithms for noninvasive, low-cost and scalable measurement of physiological (e.g., cardiac and pulmonary) vital signs is very attractive. However, diverse data representing a range of environments, body motions, illumination conditions and physiological states is laborious, time consuming and expensive to obtain. Synthetic data have proven a valuable tool in several areas of machine learning, yet are not widely available for camera measurement of physiological states. Synthetic data offer "perfect" labels (e.g., without noise and with precise synchronization), labels that may not be possible to obtain otherwise (e.g., precise pixel level segmentation maps) and provide a high degree of control over variation and diversity in the dataset. We present SCAMPS, a dataset of synthetics containing 2,800 videos (1.68M frames) with aligned cardiac and respiratory signals and facial action intensities. The RGB frames are provided alongside segmentation maps and precise descriptive statistics about the underlying waveforms, including inter-beat interval, heart rate variability, and pulse arrival time. Finally, we present baseline results training on these synthetic data and testing on real-world datasets to illustrate generalizability.

## [Robustness Disparities in Face Detection](#)

- Samuel Dooley · George Z Wei · Tom Goldstein · John Dickerson
- abstract@[open-review](#): Facial analysis systems have been deployed by large companies and critiqued by scholars and activists for the past decade. Many existing algorithmic audits examine the performance of these systems on later stage elements of facial analysis systems like facial recognition and age, emotion, or gender prediction; however, a core component to these systems has been vastly understudied from a fairness perspective: face detection. Since face detection is a pre-requisite step in facial analysis systems, the bias we observe in face detection will flow downstream to the other components like facial recognition and emotion prediction. Additionally, no prior work has focused on the robustness of these systems under various perturbations and corruptions, which leaves open the question of how various people are impacted by these phenomena. We present the first of its kind detailed benchmark of face detection systems, specifically examining the robustness to noise of commercial and academic models. We use both standard and recently released academic facial datasets to quantitatively analyze trends in face detection robustness. Across all the datasets and systems, we generally find that photos of individuals who are \emph{masculine presenting}, \emph{older}, of \emph{darker skin type}, or have \emph{dim lighting} are more susceptible to errors than their counterparts in other identities.

## [Enabling Detailed Action Recognition Evaluation Through Video Dataset Augmentation](#)

- Jihoon Chung · Yu Wu · Olga Russakovsky

- abstract@[open-review](#): It is well-known in the video understanding community that human action recognition models suffer from background bias, i.e., over-relying on scene cues in making their predictions. However, it is difficult to quantify this effect using existing evaluation frameworks. We introduce the Human-centric Analysis Toolkit (HAT), which enables evaluation of learned background bias without the need for new manual video annotation. It does so by automatically generating synthetically manipulated videos and leveraging the recent advances in image segmentation and video inpainting. Using HAT we perform an extensive analysis of 74 action recognition models trained on the Kinetics dataset. We confirm that all these models focus more on the scene background than on the human motion; further, we demonstrate that certain model design decisions (such as training with fewer frames per video or using dense as opposed to uniform temporal sampling) appear to worsen the background bias. We open-source HAT to enable the community to design more robust and generalizable human action recognition models.

## [CGLB: Benchmark Tasks for Continual Graph Learning](#)

- Xikun Zhang · Dongjin Song · Dacheng Tao
- abstract@[open-review](#): Continual learning on graph data, which aims to accommodate new tasks over newly emerged graph data while maintaining the model performance over existing tasks, is attracting increasing attention from the community. Unlike continual learning on Euclidean data (\$\text{e.g.}\}\$, images, texts, etc.) that has established benchmarks and unified experimental settings, benchmark tasks are rare for Continual Graph Learning (CGL). Moreover, due to the variety of graph data and its complex topological structures, existing works adopt different protocols to configure datasets and experimental settings. This creates a great obstacle to compare different techniques and thus hinders the development of CGL. To this end, we systematically study the task configurations in different application scenarios and develop a comprehensive Continual Graph Learning Benchmark (CGLB) curated from different public datasets. Specifically, CGLB contains both node-level and graph-level continual graph learning tasks under task-incremental (currently widely adopted) and class-incremental (more practical, challenging, yet underexplored) settings, as well as a toolkit for training, evaluating, and visualizing different CGL methods. Within CGLB, we also systematically explain the difference among these task configurations by comparing them to classical continual learning settings. Finally, we comprehensively compare state-of-the-art baselines on CGLB to investigate their effectiveness. Given CGLB and the developed toolkit, the barrier to exploring CGL has been greatly lowered and researchers can focus more on the model development without worrying about tedious work on pre-processing of datasets or encountering unseen pitfalls. The benchmark and the toolkit are available through <https://github.com/QueuQ/CGLB>.

## [Multilingual Abusive Comment Detection at Scale for Indic Languages](#)

- Vikram Gupta · Sumegh Roychowdhury · Mithun Das · Somnath Banerjee · Punyajoy Saha · Binny Mathew · hastagiri prakash vanchinathan · Animesh Mukherjee
- abstract@[open-review](#): Social media platforms were conceived to act as online town squares' where people could get together, share information and communicate with each other peacefully. However, harmful content borne out of bad actors are constantly plaguing these platforms slowly converting them into 'mosh pits' where the bad actors take the liberty to extensively abuse various marginalised groups. Accurate and timely detection of abusive content on social media platforms is therefore very important for facilitating safe interactions between users. However, due to the small scale and sparse linguistic coverage of Indic abusive speech datasets, development of such algorithms for Indic social media users (one-sixth of global population) is severely impeded. To facilitate and encourage research in this important direction, we contribute for the first time MACD - a large-scale (150K), human-annotated, multilingual (5 languages), balanced (49% abusive content) and diverse (70K users) abuse detection dataset of user comments, sourced from a popular social media platform - ShareChat. We also release AbuseXLMR, an abusive content detection model pretrained on large number of social media comments in 15+ Indic languages which outperforms XLM-R and MuRIL on multiple Indic datasets. Along with the annotations, we also release the mapping between comment, post and user id's to facilitate modelling the relationship between them. We share competitive monolingual, cross-lingual and few-shot baselines so that MACD can be used as a dataset benchmark for future research.

## [Towards Better Evaluation for Dynamic Link Prediction](#)

- Farimah Poursafaei · Shenyang Huang · Kellin Pelrine · Reihaneh Rabbany
- abstract@[open-review](#): Despite the prevalence of recent success in learning from static graphs, learning from time-evolving graphs remains an open challenge. In this work, we design new, more stringent evaluation procedures for link prediction specific to dynamic graphs, which reflect real-world considerations, to better compare the strengths and weaknesses of methods. First, we create two visualization techniques to understand the reoccurring patterns of edges over time and show that many edges reoccur at later time steps. Based on this observation, we propose a pure memorization-based baseline called EdgeBank. EdgeBank achieves surprisingly strong performance across multiple settings which highlights that the negative edges used in the current evaluation are easy. To sample more challenging negative edges, we introduce two novel negative sampling strategies that improve robustness and better match real-world applications. Lastly, we introduce six new dynamic graph datasets from a diverse set of domains missing from current benchmarks, providing new challenges and opportunities for future research. Our code repository is accessible at <https://github.com/fpour/DGB.git>.

## [CLiMB: A Continual Learning Benchmark for Vision-and-Language Tasks](#)

- Tejas Srinivasan · Ting-Yun Chang · Leticia Pinto Alva · Georgios Chochlakis · Mohammad Rostami · Jesse Thomason
- abstract@[open-review](#): Current state-of-the-art vision-and-language models are evaluated on tasks either individually or in a multi-task setting, overlooking the challenges of continually learning (CL) tasks as they arrive. Existing CL benchmarks have facilitated research on task adaptation and mitigating "catastrophic forgetting", but are limited to vision-only and language-only tasks. We present CLiMB, a benchmark to study the challenge of learning multimodal tasks in a CL setting, and to systematically evaluate how upstream continual learning can rapidly generalize to new multimodal and unimodal tasks. CLiMB includes implementations of several CL algorithms and a modified Vision-Language Transformer (ViLT) model that can be deployed on both multimodal and unimodal tasks. We find that common CL methods can help mitigate forgetting during multimodal task learning, but do not enable cross-task knowledge transfer. We envision that CLiMB will facilitate research on a new class of CL algorithms for this challenging multimodal setting.

## [AnimeRun: 2D Animation Visual Correspondence from Open Source 3D Movies](#)

- Li Siyao · Yuhang Li · Bo Li · Chao Dong · Ziwei Liu · Chen Change Loy
- abstract@[open-review](#): Visual correspondence of 2D animation is the core of many applications and deserves careful study. Existing correspondence datasets for 2D cartoon suffer from simple frame composition and monotonic movements, making them insufficient to simulate real animations. In this work, we present a new 2D animation visual correspondence dataset, AnimeRun, by converting open source 3D movies to full scenes in 2D style, including simultaneous moving background and interactions of multiple subjects. Statistics show that our proposed dataset not only resembles real anime more in image composition, but also possesses richer and more complex motion patterns compared to existing datasets. With this dataset, we establish a comprehensive benchmark by evaluating several existing optical flow and segment matching methods, and analyze shortcomings of these methods on animation data. Data are available at <https://lisiyao21.github.io/projects/AnimeRun>.

## [FLAIR: Federated Learning Annotated Image Repository](#)

- Congzheng Song · Filip Granqvist · Kunal Talwar
- abstract@[open-review](#): Cross-device federated learning is an emerging machine learning (ML) paradigm where a large population of devices collectively train an ML model while the data remains on the devices. This research field has a unique set of practical challenges, and to systematically make advances,

new datasets curated to be compatible with this paradigm are needed. Existing federated learning benchmarks in the image domain do not accurately capture the scale and heterogeneity of many real-world use cases. We introduce FLAIR, a challenging large-scale annotated image dataset for multi-label classification suitable for federated learning. FLAIR has 429,078 images from 51,414 Flickr users and captures many of the intricacies typically encountered in federated learning, such as heterogeneous user data and a long-tailed label distribution. We implement multiple baselines in different learning setups for different tasks on this dataset. We believe FLAIR can serve as a challenging benchmark for advancing the state-of-the-art in federated learning. Dataset access and the code for the benchmark are available at <https://github.com/apple/ml-flair>.

## [LAION-5B: An open large-scale dataset for training next generation image-text models](#)

- Christoph Schuhmann · Romain Beaumont · Richard Vencu · Cade Gordon · Ross Wightman · Mehdi Cherti · Theo Coombes · Aarush Katta · Clayton Mullis · Mitchell Wortsman · Patrick Schramowski · Srivatsa Kundurthy · Katherine Crowson · Ludwig Schmidt · Robert Kaczmarczyk · Jenia Jitsev
- abstract@[open-review](#): Groundbreaking language-vision architectures like CLIP and DALL-E proved the utility of training on large amounts of noisy image-text data, without relying on expensive accurate labels used in standard vision unimodal supervised learning. The resulting models showed capabilities of strong text-guided image generation and transfer to downstream tasks, while performing remarkably at zero-shot classification with noteworthy out-of-distribution robustness. Since then, large-scale language-vision models like ALIGN, BASIC, GLIDE, Flamingo and Imagen made further improvements. Studying the training and capabilities of such models requires datasets containing billions of image-text pairs. Until now, no datasets of this size have been made openly available for the broader research community. To address this problem and democratize research on large-scale multi-modal models, we present LAION-5B - a dataset consisting of 5.85 billion CLIP-filtered image-text pairs, of which 2.32B contain English language. We show successful replication and fine-tuning of foundational models like CLIP, GLIDE and Stable Diffusion using the dataset, and discuss further experiments enabled with an openly available dataset of this scale. Additionally we provide several nearest neighbor indices, an improved web-interface for dataset exploration and subset generation, and detection scores for watermark, NSFW, and toxic content detection.

## [OpenOOD: Benchmarking Generalized Out-of-Distribution Detection](#)

- Jingkang Yang · Pengyun Wang · Dejian Zou · Zitang Zhou · Kunyuan Ding · WENXUAN PENG · Haoqi Wang · Guangyao Chen · Bo Li · Yiyou Sun · Xuefeng Du · Kaiyang Zhou · Wayne Zhang · Dan Hendrycks · Yixuan Li · Ziwei Liu
- abstract@[open-review](#): Out-of-distribution (OOD) detection is vital to safety-critical machine learning applications and has thus been extensively studied, with a plethora of methods developed in the literature. However, the field currently lacks a unified, strictly formulated, and comprehensive benchmark, which often results in unfair comparisons and inconclusive results. From the problem setting perspective, OOD detection is closely related to neighboring fields including anomaly detection (AD), open set recognition (OSR), and model uncertainty, since methods developed for one domain are often applicable to each other. To help the community to improve the evaluation and advance, we build a unified, well-structured codebase called OpenOOD, which implements over 30 methods developed in relevant fields and provides a comprehensive benchmark under the recently proposed generalized OOD detection framework. With a comprehensive comparison of these methods, we are gratified that the field has progressed significantly over the past few years, where both preprocessing methods and the orthogonal post-hoc methods show strong potential.

## [Nocturne: a scalable driving benchmark for bringing multi-agent learning one step closer to the real world](#)

- Eugene Vinitsky · Nathan Lichtenstein · Xiaomeng Yang · Brandon Amos · Jakob Foerster
- abstract@[open-review](#): We introduce \textit{Nocturne}, a new 2D driving simulator for investigating multi-agent coordination under partial observability. The focus of Nocturne is to enable research into inference and theory of mind in real-world multi-agent settings without the computational overhead of computer vision and feature extraction from images. Agents in this simulator only observe an obstructed view of the scene, mimicking human visual sensing constraints. Unlike existing benchmarks that are bottlenecked by rendering human-like observations directly using a camera input, Nocturne uses efficient intersection methods to compute a vectorized set of visible features in a C++ back-end, allowing the simulator to run at \$2000+\$ steps-per-second. Using open-source trajectory and map data, we construct a simulator to load and replay arbitrary trajectories and scenes from real-world driving data. Using this environment, we benchmark reinforcement-learning and imitation-learning agents and demonstrate that the agents are quite far from human-level coordination ability and deviate significantly from the expert trajectories.

## [Dungeons and Data: A Large-Scale NetHack Dataset](#)

- Eric Hambro · Roberta Raileanu · Danielle Rothermel · Vegard Mella · Tim Rocktäschel · Heinrich Kästner · Naila Murray
- abstract@[open-review](#): Recent breakthroughs in the development of agents to solve challenging sequential decision making problems such as Go, StarCraft, or DOTA, have relied on both simulated environments and large-scale datasets. However, progress on this research has been hindered by the scarcity of open-sourced datasets and the prohibitive computational cost to work with them. Here we present the NetHack Learning Dataset (NLD), a large and highly-scalable dataset of trajectories from the popular game of NetHack, which is both extremely challenging for current methods and very fast to run. NLD consists of three parts: 10 billion state transitions from 1.5 million human trajectories collected on the NAO public NetHack server from 2009 to 2020; 3 billion state-action-score transitions from 100,000 trajectories collected from the symbolic bot winner of the NetHack Challenge 2021; and, accompanying code for users to record, load and stream any collection of such trajectories in a highly compressed form. We evaluate a wide range of existing algorithms for learning from demonstrations, showing that significant research advances are needed to fully leverage large-scale datasets for challenging sequential decision making tasks.

## [TempEL: Linking Dynamically Evolving and Newly Emerging Entities](#)

- Klim Zaporozets · Lucie-Aimée Kaffee · Johannes Deleu · Thomas Demeester · Chris Develder · Isabelle Augenstein
- abstract@[open-review](#): In our continuously evolving world, entities change over time and new, previously non-existing or unknown, entities appear. We study how this evolutionary scenario impacts the performance on a well-established entity linking (EL) task. For that study, we introduce TempEL, an entity linking dataset that consists of time-stratified English Wikipedia snapshots from 2013 to 2022, from which we collect both anchor mentions of entities, and these target entities' descriptions. By capturing such temporal aspects, our newly introduced TempEL resource contrasts with currently existing entity linking datasets, which are composed of fixed mentions linked to a single static version of a target Knowledge Base (e.g., Wikipedia 2010 for CoNLL-AIDA). Indeed, for each of our collected temporal snapshots, TempEL contains links to entities that are continual, i.e., occur in all of the years, as well as completely new entities that appear for the first time at some point. Thus, we enable to quantify the performance of current state-of-the-art EL models for: (i) entities that are subject to changes over time in their Knowledge Base descriptions as well as their mentions' contexts, and (ii) newly created entities that were previously non-existing (e.g., at the time the EL model was trained). Our experimental results show that in terms of temporal performance degradation, (i) continual entities suffer a decrease of up to 3.1% EL accuracy, while (ii) for new entities this accuracy drop is up to 17.9%. This highlights the challenge of the introduced TempEL dataset and opens new research prospects in the area of time-evolving entity disambiguation.

## [NeoRL: A Near Real-World Benchmark for Offline Reinforcement Learning](#)

- Rong-Jun Qin · Xingyuan Zhang · Songyi Gao · Xiong-Hui Chen · Zewen Li · Weinan Zhang · Yang Yu
- abstract@[open-review](#): Offline reinforcement learning (RL) aims at learning effective policies from historical data without extra environment interactions. During our experience of applying offline RL, we noticed that previous offline RL benchmarks commonly involve significant reality gaps, which we have

identified include rich and overly exploratory datasets, degraded baseline, and missing policy validation. In many real-world situations, to ensure system safety, running an overly exploratory policy to collect various data is prohibited, thus only a narrow data distribution is available. The resulting policy is regarded as effective if it is better than the working behavior policy; the policy model can be deployed only if it has been well validated, rather than accomplished the training. In this paper, we present a Near real-world offline RL benchmark, named NeoRL, to reflect these properties. NeoRL datasets are collected with a more conservative strategy. Moreover, NeoRL contains the offline training and offline validation pipeline before the online test, corresponding to real-world situations. We then evaluate recent state-of-the-art offline RL algorithms in NeoRL. The empirical results demonstrate that some offline RL algorithms are less competitive to the behavior cloning and the deterministic behavior policy, implying that they could be less effective in real-world tasks than in the previous benchmarks. We also disclose that current offline policy evaluation methods could hardly select the best policy. We hope this work will shed some light on future research and deploying RL in real-world systems.

## [BigBio: A Framework for Data-Centric Biomedical Natural Language Processing](#)

- Jason Fries · Leon Weber · Natasha Seelam · Gabriel Altay · Debajyoti Datta · Samuele Garda · Sunny Kang · Rosaline Su · Wojciech Kusa · Samuel Cahyawijaya · Fabio Barth · Simon Ott · Matthias Samwald · Stephen Bach · Stella Biderman · Mario Sanger · Bo Wang · Alison Callahan · Daniel León Periñán · Théo Gigant · Patrick Haller · Jenny Chim · Jose Posada · John Giorgi · Karthik Rangasai · Sivaraman · Marc Pámes · Marianna Nezhurina · Robert Martin · Michael Cullan · Moritz Freidank · Nathan Dahlberg · Shubhanshu Mishra · Shamik Bose · Nicholas Broad · Yanis Labrak · Shlok Deshmukh · Sid Kiblawi · Ayush Singh · Minh Chien Vu · Trishala Neeraj · Jonas Golde · Albert Villanova del Moral · Benjamin Beilharz
- abstract@[open-review](#): Training and evaluating language models increasingly requires the construction of meta-datasets -- diverse collections of curated data with clear provenance. Natural language prompting has recently lead to improved zero-shot generalization by transforming existing, supervised datasets into a variety of novel instruction tuning tasks, highlighting the benefits of meta-dataset curation. While successful in general-domain text, translating these data-centric approaches to biomedical language modeling remains challenging, as labeled biomedical datasets are significantly underrepresented in popular data hubs. To address this challenge, we introduce BigBio a community library of 126+ biomedical NLP datasets, currently covering 13 task categories and 10+ languages. BigBio facilitates reproducible meta-dataset curation via programmatic access to datasets and their metadata, and is compatible with current platforms for prompt engineering and end-to-end few/zero shot language model evaluation. We discuss our process for task schema harmonization, data auditing, contribution guidelines, and outline two illustrative use cases: zero-shot evaluation of biomedical prompts and large-scale, multi-task learning. BigBio is an ongoing community effort and is available at <https://github.com/bigscience-workshop/biomedical>

## [HAPI: A Large-scale Longitudinal Dataset of Commercial ML API Predictions](#)

- Lingjiao Chen · Zhihua Jin · Evan Sabri Eyuboglu · Christopher Răducanu · Matei Zaharia · James Zou
- abstract@[open-review](#): Commercial ML APIs offered by providers such as Google, Amazon and Microsoft have dramatically simplified ML adoptions in many applications. Numerous companies and academics pay to use ML APIs for tasks such as object detection, OCR and sentiment analysis. Different ML APIs tackling the same task can have very heterogeneous performances. Moreover, the ML models underlying the APIs also evolve over time. As ML APIs rapidly become a valuable marketplace and an integral part of analytics, it is critical to systematically study and compare different APIs with each other and to characterize how individual APIs change over time. However, this practically important topic is currently underexplored due to the lack of data. In this paper, we present HAPI (History of APIs), a longitudinal dataset of 1,761,417 instances of commercial ML API applications (involving APIs from Amazon, Google, IBM, Microsoft and other providers) across diverse tasks including image tagging, speech recognition, and text mining from 2020 to 2022. Each instance consists of a query input for an API (e.g., an image or text) along with the API's output prediction/annotation and confidence scores. HAPI is the first large-scale dataset of ML API usages and is a unique resource for studying ML as-a-service (MLaaS). As examples of the types of analyses that HAPI enables, we show that ML APIs' performance changes substantially over time—"several APIs' accuracies dropped on specific benchmark datasets. Even when the API's aggregate performance stays steady, its error modes can shift across different subtypes of data between 2020 and 2022. Such changes can substantially impact the entire analytics pipelines that use some ML API as a component. We further use HAPI to study commercial APIs' performance disparities across demographic subgroups over time. HAPI can stimulate more research in the growing field of MLaaS.

## [DART: Articulated Hand Model with Diverse Accessories and Rich Textures](#)

- Diheng Gao · Yuliang Xiu · Kailin Li · Lixin Yang · Feng Wang · Peng Zhang · Bang Zhang · Cewu Lu · Ping Tan
- abstract@[open-review](#): Hand, the bearer of human productivity and intelligence, is receiving much attention due to the recent fever of digital twins. Among different hand morphable models, MANO has been widely used in vision and graphics community. However, MANO disregards textures and accessories, which largely limits its power to synthesize photorealistic hand data. In this paper, we extend MANO with Diverse Accessories and Rich Textures, namely DART. DART is composed of 50 daily 3D accessories which varies in appearance and shape, and 325 hand-crafted 2D texture maps covers different kinds of blemishes or make-ups. Unity GUI is also provided to generate synthetic hand data with user-defined settings, e.g., pose, camera, background, lighting, textures, and accessories. Finally, we release DARTset, which contains large-scale (800K), high-fidelity synthetic hand images, paired with perfect-aligned 3D labels. Experiments demonstrate its superiority in diversity. As a complement to existing hand datasets, DARTset boosts the generalization in both hand pose estimation and mesh recovery tasks. Raw ingredients (textures, accessories), Unity GUI, source code and DARTset are publicly available at [dart2022.github.io](https://dart2022.github.io).

## [DGraph: A Large-Scale Financial Dataset for Graph Anomaly Detection](#)

- Xuanwen Huang · Yang Yang · Yang Wang · Chunping Wang · Zhisheng Zhang · Jiarong Xu · Lei Chen · Michalis Vazirgiannis
- abstract@[open-review](#): Graph Anomaly Detection (GAD) has recently become a hot research spot due to its practicability and theoretical value. Since GAD emphasizes the application and the rarity of anomalous samples, enriching the varieties of its datasets is fundamental. Thus, this paper present DGraph, a real-world dynamic graph in the finance domain. DGraph overcomes many limitations of current GAD datasets. It contains about 3M nodes, 4M dynamic edges, and 1M ground-truth nodes. We provide a comprehensive observation of DGraph, revealing that anomalous nodes and normal nodes generally have different structures, neighbor distribution, and temporal dynamics. Moreover, it suggests that 2M background nodes are also essential for detecting fraudsters. Furthermore, we conduct extensive experiments on DGraph. Observation and experiments demonstrate that DGraph is propulsive to advance GAD research and enable in-depth exploration of anomalous nodes.

## [Understanding Aesthetics with Language: A Photo Critique Dataset for Aesthetic Assessment](#)

- Daniel Vera Nieto · Luigi Celona · Clara Fernandez Labrador
- abstract@[open-review](#): Computational inference of aesthetics is an ill-defined task due to its subjective nature. Many datasets have been proposed to tackle the problem by providing pairs of images and aesthetic scores based on human ratings. However, humans are better at expressing their opinion, taste, and emotions by means of language rather than summarizing them in a single number. In fact, photo critiques provide much richer information as they reveal how and why users rate the aesthetics of visual stimuli. In this regard, we propose the Reddit Photo Critique Dataset (RPCD), which contains tuples of image and photo critiques. RPCD consists of 74K images and 220K comments and is collected from a Reddit community used by hobbyists and professional photographers to improve their photography skills by leveraging constructive community feedback. The proposed dataset differs from previous aesthetics datasets mainly in three aspects, namely (i) the large scale of the dataset and the extension of the comments criticizing different aspects of the image, (ii) it contains mostly UltraHD images, and (iii) it can easily be extended to new data as it is collected through an automatic pipeline. To the

best of our knowledge, in this work, we propose the first attempt to estimate the aesthetic quality of visual stimuli from the critiques. To this end, we exploit the polarity of the sentiment of criticism as an indicator of aesthetic judgment. We demonstrate how sentiment polarity correlates positively with the aesthetic judgment available for two aesthetic assessment benchmarks. Finally, we experiment with several models by using the sentiment scores as a target for ranking images. Dataset and baselines are available <https://github.com/mediatechnologycenter/aestheval>.

## [PulseImpute: A Novel Benchmark Task for Pulsative Physiological Signal Imputation](#)

- Maxwell Xu · Alexander Moreno · Supriya Nagesh · Varol Aydemir · David Wetter · Santosh Kumar · James Rehg
- abstract@[open-review](#): The promise of Mobile Health (mHealth) is the ability to use wearable sensors to monitor participant physiology at high frequencies during daily life to enable temporally-precise health interventions. However, a major challenge is frequent missing data. Despite a rich imputation literature, existing techniques are ineffective for the pulsative signals which comprise many mHealth applications, and a lack of available datasets has stymied progress. We address this gap with PulseImpute, the first large-scale pulsative signal imputation challenge which includes realistic mHealth missingness models, an extensive set of baselines, and clinically-relevant downstream tasks. Our baseline models include a novel transformer-based architecture designed to exploit the structure of pulsative signals. We hope that PulseImpute will enable the ML community to tackle this significant and challenging task.

## [Meta-Album: Multi-domain Meta-Dataset for Few-Shot Image Classification](#)

- Ihsan Ullah · Dustin Carrión-Ojeda · Sergio Escalera · Isabelle Guyon · Mike Huisman · Felix Mohr · Jan N. van Rijn · Haozhe Sun · Joaquin Vanschoren · Phan Anh Vu
- abstract@[open-review](#): We introduce Meta-Album, an image classification meta-dataset designed to facilitate few-shot learning, transfer learning, meta-learning, among other tasks. It includes 40 open datasets, each having at least 20 classes with 40 examples per class, with verified licences. They stem from diverse domains, such as ecology (fauna and flora), manufacturing (textures, vehicles), human actions, and optical character recognition, featuring various image scales (microscopic, human scales, remote sensing). All datasets are preprocessed, annotated, and formatted uniformly, and come in 3 versions (Micro \$\subset\$ Mini \$\subset\$ Extended) to match users' computational resources. We showcase the utility of the first 30 datasets on few-shot learning problems. The other 10 will be released shortly after. Meta-Album is already more diverse and larger (in number of datasets) than similar efforts, and we are committed to keep enlarging it via a series of competitions. As competitions terminate, their test data are released, thus creating a rolling benchmark, available through OpenML.org. Our website <https://meta-album.github.io/> contains the source code of challenge winning methods, baseline methods, data loaders, and instructions for contributing either new datasets or algorithms to our expandable meta-dataset.

## [On Reinforcement Learning and Distribution Matching for Fine-Tuning Language Models with no Catastrophic Forgetting](#)

- Tomasz Korbak · Hady Elsahar · Germán Kruszewski · Marc Dymetman
- abstract@[open-review](#): The availability of large pre-trained models is changing the landscape of Machine Learning research and practice, moving from a "training from scratch" to a "fine-tuning" paradigm. While in some applications the goal is to "nudge" the pre-trained distribution towards preferred outputs, in others it is to steer it towards a different distribution over the sample space. Two main paradigms have emerged to tackle this challenge: Reward Maximization (RM) and, more recently, Distribution Matching (DM). RM applies standard Reinforcement Learning (RL) techniques, such as Policy Gradients, to gradually increase the reward signal. DM prescribes to first make explicit the target distribution that the model is fine-tuned to approximate. Here we explore the theoretical connections between the two paradigms and show that methods such as KL-control developed in the RM paradigm can also be construed as belonging to DM. We further observe that while DM differs from RM, it can suffer from similar training difficulties, such as high gradient variance. We leverage connections between the two paradigms to import the concept of baseline into DM methods. We empirically validate the benefits of adding a baseline on an array of controllable language generation tasks such as constraining topic, sentiment, and gender distributions in texts sampled from a language model. We observe superior performance in terms of constraint satisfaction, stability, and sample efficiency.

## [Adaptive Interest for Emphatic Reinforcement Learning](#)

- Martin Klissarov · Rasool Fakoor · Jonas Mueller · Kavosh Asadi · Taesup Kim · Alexander Smola
- abstract@[open-review](#): Emphatic algorithms have shown great promise in stabilizing and improving reinforcement learning by selectively emphasizing the update rule. Although the emphasis fundamentally depends on an interest function which defines the intrinsic importance of each state, most approaches simply adopt a uniform interest over all states (except where a hand-designed interest is possible based on domain knowledge). In this paper, we investigate adaptive methods that allow the interest function to dynamically vary over states and iterations. In particular, we leverage meta-gradients to automatically discover online an interest function that would accelerate the agent's learning process. Empirical evaluations on a wide range of environments show that adapting the interest is key to provide significant gains. Qualitative analysis indicates that the learned interest function emphasizes states of particular importance, such as bottlenecks, which can be especially useful in a transfer learning setting.

## [Hilbert Distillation for Cross-Dimensionality Networks](#)

- Dian Qin · Haishuai Wang · Zhe Liu · HONGJIA XU · Sheng Zhou · Jiajun Bu
- abstract@[open-review](#): 3D convolutional neural networks have revealed superior performance in processing volumetric data such as video and medical imaging. However, the competitive performance by leveraging 3D networks results in huge computational costs, which are far beyond that of 2D networks. In this paper, we propose a novel Hilbert curve-based cross-dimensionality distillation approach that facilitates the knowledge of 3D networks to improve the performance of 2D networks. The proposed Hilbert Distillation (HD) method preserves the structural information via the Hilbert curve, which maps high-dimensional ( $>=2$ ) representations to one-dimensional continuous space-filling curves. Since the distilled 2D networks are supervised by the curves converted from dimensionally heterogeneous 3D features, the 2D networks are given an informative view in terms of learning structural information embedded in well-trained high-dimensional representations. We further propose a Variable-length Hilbert Distillation (VHD) method to dynamically shorten the walking stride of the Hilbert curve in activation feature areas and lengthen the stride in context feature areas, forcing the 2D networks to pay more attention to learning from activation features. The proposed algorithm outperforms the current state-of-the-art distillation techniques adapted to cross-dimensionality distillation on two classification tasks. Moreover, the distilled 2D networks by the proposed method achieve competitive performance with the original 3D networks, indicating the lightweight distilled 2D networks could potentially be the substitution of cumbersome 3D networks in the real-world scenario.

## [Distributionally Adaptive Meta Reinforcement Learning](#)

- Anurag Ajay · Dibya Ghosh · Sergey Levine · Pulkit Agrawal · Abhishek Gupta
- abstract@[open-review](#): Meta-reinforcement learning algorithms provide a data-driven way to acquire learning algorithms that quickly adapt to many tasks with varying rewards or dynamics functions. However, learned meta-policies are often effective only on the exact task distribution on which the policy was trained, and struggle in the presence of distribution shift of test-time rewards or transition dynamics. In this work, we develop a framework for meta-RL algorithms that are able to behave appropriately under test-time distribution shifts in the space of tasks. Our framework centers on an adaptive approach to distributional robustness, in which we train a population of meta-agents to be robust to varying levels of distribution shift, so that when evaluated on a (potentially shifted) test-time distribution of tasks, we can adaptively choose the most appropriate meta-agent to follow. We formally show how this framework allows for improved regret under distribution shift, and empirically show its efficacy on simulated robotics problems under a wide range of distribution shifts.

## [Simplified Graph Convolution with Heterophily](#)

- Sudhanshu Chanpuriya · Cameron Musco
- abstract@[open-review](#): Recent work has shown that a simple, fast method called Simple Graph Convolution (SGC) (Wu et al., 2019), which eschews deep learning, is competitive with deep methods like graph convolutional networks (GCNs) (Kipf & Welling, 2017) in common graph machine learning benchmarks. The use of graph data in SGC implicitly assumes the common but not universal graph characteristic of homophily, wherein nodes link to nodes which are similar. Here we confirm that SGC is indeed ineffective for heterophilous (i.e., non-homophilous) graphs via experiments on synthetic and real-world datasets. We propose Adaptive Simple Graph Convolution (ASGC), which we show can adapt to both homophilous and heterophilous graph structure. Like SGC, ASGC is not a deep model, and hence is fast, scalable, and interpretable; further, we can prove performance guarantees on natural synthetic data models. Empirically, ASGC is often competitive with recent deep models at node classification on a benchmark of real-world datasets. The SGC paper questioned whether the complexity of graph neural networks is warranted for common graph problems involving homophilous networks; our results similarly suggest that, while deep learning often achieves the highest performance, heterophilous structure alone does not necessitate these more involved methods.

## [Accelerating Certified Robustness Training via Knowledge Transfer](#)

- Pratik Vaishnavi · Kevin Eykholt · Amir Rahmati
- abstract@[open-review](#): Training deep neural network classifiers that are certifiably robust against adversarial attacks is critical to ensuring the security and reliability of AI-controlled systems. Although numerous state-of-the-art certified training methods have been developed, they are computationally expensive and scale poorly with respect to both dataset and network complexity. Widespread usage of certified training is further hindered by the fact that periodic retraining is necessary to incorporate new data and network improvements. In this paper, we propose Certified Robustness Transfer (CRT), a general-purpose framework for reducing the computational overhead of any certifiably robust training method through knowledge transfer. Given a robust teacher, our framework uses a novel training loss to transfer the teacher's robustness to the student. We provide theoretical and empirical validation of CRT. Our experiments on CIFAR-10 show that CRT speeds up certified robustness training by \$8 \times\$ on average across three different architecture generations, while achieving comparable robustness to state-of-the-art methods. We also show that CRT can scale to large-scale datasets like ImageNet.

## [Empirical Gateaux Derivatives for Causal Inference](#)

- Michael Jordan · Yixin Wang · Angela Zhou
- abstract@[open-review](#): We study a constructive procedure that approximates Gateaux derivatives for statistical functionals by finite-differencing, with attention to causal inference functionals. We focus on the case where probability distributions are not known a priori but need also to be estimated from data, leading to empirical Gateaux derivatives, and study relationships between empirical, numerical, and analytical Gateaux derivatives. Starting with a case study of counterfactual mean estimation, we verify the exact relationship between finite-differences and the analytical Gateaux derivative. We then derive requirements on the rates of numerical approximation in perturbation and smoothing that preserve statistical benefits. We study more complicated functionals such as dynamic treatment regimes and the linear-programming formulation for policy optimization infinite-horizon Markov decision processes. In the case of the latter, this approach can be used to approximate bias adjustments in the presence of arbitrary constraints, illustrating the usefulness of constructive approaches for Gateaux derivatives. We find that, omitting unfavorable dimension dependence of smoothing, although rate-double robustness permits for coarser rates of perturbation size than implied by generic approximation analysis of finite-differences for the case of the counterfactual mean, this is not the case for the infinite-horizon MDP policy value.

## [Beyond Separability: Analyzing the Linear Transferability of Contrastive Representations to Related Subpopulations](#)

- Jeff Z. HaoChen · Colin Wei · Ananya Kumar · Tengyu Ma
- abstract@[open-review](#): Contrastive learning is a highly effective method for learning representations from unlabeled data. Recent works show that contrastive representations can transfer across domains, leading to simple state-of-the-art algorithms for unsupervised domain adaptation. In particular, a linear classifier trained to separate the representations on the source domain can also predict classes on the target domain accurately, even though the representations of the two domains are far from each other. We refer to this phenomenon as linear transferability. This paper analyzes when and why contrastive representations exhibit linear transferability in a general unsupervised domain adaptation setting. We prove that linear transferability can occur when data from the same class in different domains (e.g., photo dogs and cartoon dogs) are more related with each other than data from different classes in different domains (e.g., photo dogs and cartoon cats) are. Our analyses are in a realistic regime where the source and target domains can have unbounded density ratios and be weakly related, and they have distant representations across domains.

## [Efficient and Near-Optimal Smoothed Online Learning for Generalized Linear Functions](#)

- Adam Block · Max Simchowitz
- abstract@[open-review](#): Due to the drastic gap in complexity between sequential and batch statistical learning, recent work has studied a smoothed sequential learning setting, where Nature is constrained to select contexts with density bounded by  $\sigma$  with respect to a known measure  $\mu$ . Unfortunately, for some function classes, there is an exponential gap between the statistically optimal regret and that which can be achieved efficiently. In this paper, we give a computationally efficient algorithm that is the first to enjoy the statistically optimal  $\log(T/\sigma)$  regret for realizable  $K$ -wise linear classification. We extend our results to settings where the true classifier is linear in an over-parameterized polynomial featurization of the contexts, as well as to a realizable piecewise-regression setting assuming access to an appropriate ERM oracle. Somewhat surprisingly, standard disagreement-based analyses are insufficient to achieve regret logarithmic in  $\sigma$ . Instead, we develop a novel characterization of the geometry of the disagreement region induced by generalized linear classifiers. Along the way, we develop numerous technical tools of independent interest, including a general anti-concentration bound for the determinant of certain matrix averages.

## [ShapeCrafter: A Recursive Text-Conditioned 3D Shape Generation Model](#)

- Rao Fu · Xiao Zhan · YIWEN CHEN · Daniel Ritchie · Srinath Sridhar
- abstract@[open-review](#): We present ShapeCrafter, a neural network for recursive text-conditioned 3D shape generation. Existing methods to generate text-conditioned 3D shapes consume an entire text prompt to generate a 3D shape in a single step. However, humans tend to describe shapes recursively--we may start with an initial description and progressively add details based on intermediate results. To capture this recursive process, we introduce a method to generate a 3D shape distribution, conditioned on an initial phrase, that gradually evolves as more phrases are added. Since existing datasets are insufficient for training this approach, we present Text2Shape++, a large dataset of 369K shape-text pairs that supports recursive shape generation. To capture local details that are often used to refine shape descriptions, we build on top of vector-quantized deep implicit functions that generate a distribution of high-quality shapes. Results show that our method can generate shapes consistent with text descriptions, and shapes evolve gradually as more phrases are added. Our method supports shape editing, extrapolation, and can enable new applications in human-machine collaboration for creative design.

## [Trajectory Inference via Mean-field Langevin in Path Space](#)

- Stephen Zhang · Léonard Chizat · Matthieu Heitz · Geoffrey Schiebinger

- abstract@[open-review](#): Trajectory inference aims at recovering the dynamics of a population from snapshots of its temporal marginals. To solve this task, a min-entropy estimator relative to the Wiener measure in path space was introduced in [Laventan et al., 2021], and shown to consistently recover the dynamics of a large class of drift-diffusion processes from the solution of an infinite dimensional convex optimization problem. In this paper, we introduce a grid-free algorithm to compute this estimator. Our method consists in a family of point clouds (one per snapshot) coupled via Schrödinger bridges which evolve with noisy gradient descent. We study the mean-field limit of the dynamics and prove its global convergence to the desired estimator. Overall, this leads to an inference method with end-to-end theoretical guarantees that solves an interpretable model for trajectory inference. We also present how to adapt the method to deal with mass variations, a useful extension when dealing with single cell RNA-sequencing data where cells can branch and die.

## [Beyond black box densities: Parameter learning for the deviated components](#)

- Dat Do · Nhat Ho · XuanLong Nguyen
- abstract@[open-review](#): As we collect additional samples from a data population for which a known density function estimate may have been previously obtained by a black box method, the increased complexity of the data set may result in the true density being deviated from the known estimate by a mixture distribution. To model this phenomenon, we consider the deviating mixture model  $(1-\lambda)h_0 + \lambda p_i f(x|\theta_i)$ , where  $h_0$  is a known density function, while the deviated proportion  $\lambda$  and latent mixing measure  $G_i = \sum_{i=1}^k p_i \delta_{\theta_i}$  associated with the mixture distribution are unknown. Via a novel notion of distinguishability between the known density  $h_0$  and the deviated mixture distribution, we establish rates of convergence for the maximum likelihood estimates of  $\lambda$  and  $G$  under Wasserstein metric. Simulation studies are carried out to illustrate the theory.

## [Stability and Generalization of Kernel Clustering: from Single Kernel to Multiple Kernel](#)

- Weixuan Liang · Xinwang Liu · Yong Liu · sihang zhou · Jun-Jie Huang · Siwei Wang · Jiyuan Liu · Yi Zhang · En Zhu
- abstract@[open-review](#): Multiple kernel clustering (MKC) is an important research topic that has been widely studied for decades. However, current methods still face two problems: inefficient when handling out-of-sample data points and lack of theoretical study of the stability and generalization of clustering. In this paper, we propose a novel method that can efficiently compute the embedding of out-of-sample data with a solid generalization guarantee. Specifically, we approximate the eigen functions of the integral operator associated with the linear combination of base kernel functions to construct low-dimensional embeddings of out-of-sample points for efficient multiple kernel clustering. In addition, we, for the first time, theoretically study the stability of clustering algorithms and prove that the single-view version of the proposed method has uniform stability as  $\mathcal{O}(Kn^{-3/2})$  and establish an upper bound of excess risk as  $\widetilde{\mathcal{O}}(Kn^{-3/2} + n^{-1/2})$ , where  $K$  is the cluster number and  $n$  is the number of samples. We then extend the theoretical results to multiple kernel scenarios and find that the stability of MKC depends on kernel weights. As an example, we apply our method to a novel MKC algorithm termed SimpleMKM and derive the upper bound of its excess clustering risk, which is tighter than the current results. Extensive experimental results validate the effectiveness and efficiency of the proposed method.

## [Leveraging the Hints: Adaptive Bidding in Repeated First-Price Auctions](#)

- Wei Zhang · Yanjun Han · Zhengyuan Zhou · Aaron Flores · Tsachy Weissman
- abstract@[open-review](#): With the advent and increasing consolidation of e-commerce, digital advertising has very recently replaced traditional advertising as the main marketing force in the economy. In the past four years, a particularly important development in the digital advertising industry is the shift from second-price auctions to first-price auctions for online display ads. This shift immediately motivated the intellectually challenging question of how to bid in first-price auctions, because unlike in second-price auctions, bidding one's private value truthfully is no longer optimal. Following a series of recent works in this area, we consider a differentiated setup: we do not make any assumption about other bidders' maximum bid (i.e. it can be adversarial over time), and instead assume that we have access to a hint that serves as a prediction of other bidders' maximum bid, where the prediction is learned through some blackbox machine learning model. We consider two types of hints: one where a single point-prediction is available, and the other where a hint interval (representing a type of confidence region into which others' maximum bid falls) is available. We establish minimax optimal regret bounds for both cases and highlight the quantitatively different behavior between the two settings. We also provide improved regret bounds when the others' maximum bid exhibits the further structure of sparsity. Finally, we complement the theoretical results with demonstrations using real bidding data.

## [Boosting Barely Robust Learners: A New Perspective on Adversarial Robustness](#)

- Avrim Blum · Omar Montasser · Greg Shakhnarovich · Hongyang Zhang
- abstract@[open-review](#): We present an oracle-efficient algorithm for boosting the adversarial robustness of barely robust learners. Barely robust learning algorithms learn predictors that are adversarially robust only on a small fraction  $\beta$  of the data distribution. Our proposed notion of barely robust learning requires robustness with respect to a "larger" perturbation set; which we show is necessary for strongly robust learning, and that weaker relaxations are not sufficient for strongly robust learning. Our results reveal a qualitative and quantitative equivalence between two seemingly unrelated problems: strongly robust learning and barely robust learning.

## [Robust Anytime Learning of Markov Decision Processes](#)

- Marnix Suijen · Thiago D. Simão · Nils Jansen · David Parker
- abstract@[open-review](#): Markov decision processes (MDPs) are formal models commonly used in sequential decision-making. MDPs capture the stochasticity that may arise, for instance, from imprecise actuators via probabilities in the transition function. However, in data-driven applications, deriving precise probabilities from (limited) data introduces statistical errors that may lead to unexpected or undesirable outcomes. Uncertain MDPs (uMDPs) do not require precise probabilities but instead use so-called uncertainty sets in the transitions, accounting for such limited data. Tools from the formal verification community efficiently compute robust policies that provably adhere to formal specifications, like safety constraints, under the worst-case instance in the uncertainty set. We continuously learn the transition probabilities of an MDP in a robust anytime-learning approach that combines a dedicated Bayesian inference scheme with the computation of robust policies. In particular, our method (1) approximates probabilities as intervals, (2) adapts to new data that may be inconsistent with an intermediate model, and (3) may be stopped at any time to compute a robust policy on the uMDP that faithfully captures the data so far. We show the effectiveness of our approach and compare it to robust policies computed on uMDPs learned by the UCRL2 reinforcement learning algorithm in an experimental evaluation on several benchmarks.

## [Subsidiary Prototype Alignment for Universal Domain Adaptation](#)

- Jogendra Nath Kundu · Suvaansh Bhambri · Akshay R Kulkarni · Hiran Sarkar · Varun Jampani · Venkatesh Babu R
- abstract@[open-review](#): Universal Domain Adaptation (UniDA) deals with the problem of knowledge transfer between two datasets with domain-shift as well as category-shift. The goal is to categorize unlabeled target samples, either into one of the "known" categories or into a single "unknown" category. A major problem in UniDA is negative transfer, i.e. misalignment of "known" and "unknown" classes. To this end, we first uncover an intriguing tradeoff between negative-transfer-risk and domain-invariance exhibited at different layers of a deep network. It turns out we can strike a balance between these two metrics at a mid-level layer. Towards designing an effective framework based on this insight, we draw motivation from Bag-of-visual-Words (BoW). Word-prototypes in a BoW-like representation of a mid-level layer would represent lower-level visual primitives that are likely to be unaffected by the category-shift in the high-level features. We develop modifications that encourage learning of word-prototypes followed by word-histogram based

classification. Following this, subsidiary prototype-space alignment (SPA) can be seen as a closed-set alignment problem, thereby avoiding negative transfer. We realize this with a novel word-histogram-related pretext task to enable closed-set SPA, operating in conjunction with goal task UniDA. We demonstrate the efficacy of our approach on top of existing UniDA techniques, yielding state-of-the-art performance across three standard UniDA and Open-Set DA object recognition benchmarks.

## [Generative multitask learning mitigates target-causing confounding](#)

- Taro Makino · Krzysztof Geras · Kyunghyun Cho
- abstract@[open-review](#): We propose a simple and scalable approach to causal representation learning for multitask learning. Our approach requires minimal modification to existing ML systems, and improves robustness to target shift. The improvement comes from mitigating unobserved confounders that cause the targets, but not the input. We refer to them as target-causing confounders. These confounders induce spurious dependencies between the input and targets. This poses a problem for the conventional approach to multitask learning, due to its assumption that the targets are conditionally independent given the input. Our proposed approach takes into account the dependencies between the targets in order to alleviate target-causing confounding. All that is required in addition to usual practice is to estimate the joint distribution of the targets to switch from discriminative to generative classification, and to predict all targets jointly. Our results on the Attributes of People and Taskonomy datasets reflect the conceptual improvement in robustness to target shift.

## [IM-Loss: Information Maximization Loss for Spiking Neural Networks](#)

- Yufei Guo · Yuanpei Chen · Liwen Zhang · Xiaode Liu · Yinglei Wang · Xuhui Huang · Zhe Ma
- abstract@[open-review](#): Spiking Neural Network (SNN), recognized as a type of biologically plausible architecture, has recently drawn much research attention. It transmits information by \$0/1\\$ spikes. This bio-mimetic mechanism of SNN demonstrates extreme energy efficiency since it avoids any multiplications on neuromorphic hardware. However, the forward-passing \$0/1\\$ spike quantization will cause information loss and accuracy degradation. To deal with this problem, the Information maximization loss (IM-Loss) that aims at maximizing the information flow in the SNN is proposed in the paper. The IM-Loss not only enhances the information expressiveness of an SNN directly but also plays a part of the role of normalization without introducing any additional operations (e.g., bias and scaling) in the inference phase. Additionally, we introduce a novel differentiable spike activity estimation, Evolutionary Surrogate Gradients (ESG) in SNNs. By appointing automatic evolvable surrogate gradients for spike activity function, ESG can ensure sufficient model updates at the beginning and accurate gradients at the end of the training, resulting in both easy convergence and high task performance. Experimental results on both popular non-spiking static and neuromorphic datasets show that the SNN models trained by our method outperform the current state-of-the-art algorithms.

## [What You See is What You Get: Distributional Generalization for Algorithm Design in Deep Learning](#)

- Bogdan Kulynych · Yao-Yuan Yang · Yaodong Yu · Jarosław Basiok · Preetum Nakkiran
- abstract@[open-review](#): We investigate and leverage a connection between Differential Privacy (DP) and the recently proposed notion of Distributional Generalization (DG). Applying this connection, we introduce new conceptual tools for designing deep-learning methods that bypass "pathologies" of standard stochastic gradient descent (SGD). First, we prove that differentially private methods satisfy a "What You See Is What You Get (WYSIWYG)" generalization guarantee: whatever a model does on its train data is almost exactly what it will do at test time. This guarantee is formally captured by distributional generalization. WYSIWYG enables principled algorithm design in deep learning by reducing generalization concerns to optimization ones: in order to mitigate unwanted behavior at test time, it is provably sufficient to mitigate this behavior on the train data. This is notably false for standard (non-DP) methods, hence this observation has applications even when privacy is not required. For example, importance sampling is known to fail for standard SGD, but we show that it has exactly the intended effect for DP-trained models. We use these insights to construct simple algorithms which are competitive with SOTA in several distributional robustness applications, and to significantly improve the privacy vs. disparate impact tradeoff of DP-SGD. Finally, we also improve on known theoretical bounds relating DP, stability, and distributional generalization.

## [Local Metric Learning for Off-Policy Evaluation in Contextual Bandits with Continuous Actions](#)

- Haanvid Lee · Jongmin Lee · Yunseon Choi · Wonseok Jeon · Byung-Jun Lee · Yung-Kyun Noh · Kee-Eung Kim
- abstract@[open-review](#): We consider local kernel metric learning for off-policy policy evaluation (OPE) of deterministic policies in contextual bandits with continuous action spaces. Our work is motivated by practical scenarios where the target policy needs to be deterministic due to domain requirements, such as prescription of treatment dosage and duration in medicine. Although importance sampling (IS) provides a basic principle for OPE, it is ill-posed for the deterministic target policy with continuous actions. Our main idea is to relax the target policy and pose the problem as kernel-based estimation, while learning the kernel metric in order to minimize the overall mean square error (MSE). We present an analytic solution for the optimal metric, based on the analysis of bias and variance. Whereas prior work has been limited to scalar action spaces or kernel bandwidth optimization, our work takes a step further being capable of vector action spaces and metric optimization. We show that our estimator is consistent, and significantly reduces MSE compared to baseline OPE methods through experiments on various domains.

## [Optimal Query Complexities for Dynamic Trace Estimation](#)

- David Woodruff · Fred Zhang · Richard Zhang
- abstract@[open-review](#): We consider the problem of minimizing the number of matrix-vector queries needed for accurate trace estimation in the dynamic setting where our underlying matrix is changing slowly, such as during an optimization process. Specifically, for any \$m\$ matrices \$\mathbf{A}\_1, \dots, \mathbf{A}\_m\$ with consecutive differences bounded in Schatten-\$1\$ norm by \$\alpha\$, we provide a novel binary tree summation procedure that simultaneously estimates all \$m\$ traces up to \$\epsilon\$ error with \$\delta\$ failure probability with an optimal query complexity of \$\tilde{O}(m \alpha \sqrt{\log(1/\delta)} / \epsilon + m \log(1/\delta))\$, improving the dependence on both \$\alpha\$ and \$\delta\$ from Dharangutte and Musco (NeurIPS, 2021). Our procedure works without additional norm bounds on \$\mathbf{A}\_i\$ and can be generalized to a bound for the \$p\$-th Schatten norm for \$p \in [1,2]\$, giving a complexity of \$\tilde{O}(m \alpha (\sqrt{\log(1/\delta)} / \epsilon)^p + m \log(1/\delta))\$. By using novel reductions to communication complexity and information-theoretic analyses of Gaussian matrices, we provide matching lower bounds for static and dynamic trace estimation in all relevant parameters, including the failure probability. Our lower bounds (1) give the first tight bounds for Hutchinson's estimator in the matrix-vector product model with Frobenius norm error (it even in the static setting), and (2) are the first unconditional lower bounds for dynamic trace estimation, resolving open questions of prior work.

## [Neural Topological Ordering for Computation Graphs](#)

- Mukul Agrani · Corrado Rainone · Yang Yang · Harris Teague · Wonseok Jeon · Roberto Bondesan · Herke van Hoof · Christopher Lott · Weiliang Zeng · Piero Zappi
- abstract@[open-review](#): Recent works on machine learning for combinatorial optimization have shown that learning based approaches can outperform heuristic methods in terms of speed and performance. In this paper, we consider the problem of finding an optimal topological order on a directed acyclic graph (DAG) with focus on the memory minimization problem which arises in compilers. We propose an end-to-end machine learning based approach for topological ordering using an encoder-decoder framework. Our encoder is a novel attention based graph neural network architecture called Topoformer which uses different topological transforms of a DAG for message passing. The node embeddings produced by the encoder are converted into node priorities which are used by the decoder to generate a probability distribution over topological orders. We train our model on a dataset

of synthetically generated graphs called layered graphs. We show that our model outperforms, or is on-par, with several topological ordering baselines while being significantly faster on synthetic graphs with up to 2k nodes. We also train and test our model on a set of real-world computation graphs, showing performance improvements.

## [Near-Optimal Sample Complexity Bounds for Constrained MDPs](#)

- Sharan Vaswani · Lin Yang · Csaba Szepesvari
- abstract@[open-review](#): In contrast to the advances in characterizing the sample complexity for solving Markov decision processes (MDPs), the optimal statistical complexity for solving constrained MDPs (CMDPs) remains unknown. We resolve this question by providing minimax upper and lower bounds on the sample complexity for learning near-optimal policies in a discounted CMDP with access to a generative model (simulator). In particular, we design a model-based algorithm that addresses two settings: (i) relaxed feasibility, where small constraint violations are allowed, and (ii) strict feasibility, where the output policy is required to satisfy the constraint. For (i), we prove that our algorithm returns an  $\$\\epsilon$ -optimal policy with probability  $1 - \\delta$ , by making  $\\tilde{O}(\\frac{S A \\log(1/\\delta)}{(1 - \\gamma)^3 \\epsilon^2})$  queries to the generative model, thus matching the sample-complexity for unconstrained MDPs. For (ii), we show that the algorithm's sample complexity is upper-bounded by  $\\tilde{O}(\\frac{S A \\log(1/\\delta)}{(1 - \\gamma)^5 \\epsilon^2 \\zeta^2})$  where  $\\zeta$  is the problem-dependent Slater constant that characterizes the size of the feasible region. Finally, we prove a matching lower-bound for the strict feasibility setting, thus obtaining the first near minimax optimal bounds for discounted CMDPs. Our results show that learning CMDPs is as easy as MDPs when small constraint violations are allowed, but inherently more difficult when we demand zero constraint violation.

## [When Adversarial Training Meets Vision Transformers](#)

- Yichuan Mo · Dongxian Wu · Yifei Wang · Yiwen Guo · Yisen Wang
- abstract@[open-review](#): Vision Transformers (ViTs) have recently achieved competitive performance in broad vision tasks. Unfortunately, on popular threat models, naturally trained ViTs are proven to provide no more adversarial robustness than convolutional neural networks (CNNs). Adversarial training is still required by ViTs to defend against such adversarial attacks. This paper provides a comprehensive evaluation of various training techniques across several datasets, thus bringing the first implementation benchmark for adversarial training of ViTs. In addition, regarding ViT as a new type of model architecture, we further investigate its adversarial robustness from the perspective of its unique architectural components. We find, when we randomly mask the perturbation on some of the patches or the gradient from some attention blocks during adversarial training, the adversarial robustness of ViTs can be further improved. Our work may potentially open up a line of work to explore the architectural information inside new models like ViTs.

## [Nonlinear MCMC for Bayesian Machine Learning](#)

- James Vuckovic
- abstract@[open-review](#): We explore the application of a nonlinear MCMC technique first introduced in [1] to problems in Bayesian machine learning. We provide a convergence guarantee in total variation that uses novel results for long-time convergence and large-particle ("propagation of chaos") convergence. We apply this nonlinear MCMC technique to sampling problems including a Bayesian neural network on CIFAR10.

## [Robust Neural Posterior Estimation and Statistical Model Criticism](#)

- Daniel Ward · Patrick Cannon · Mark Beaumont · Matteo Fasiolo · Sebastian Schmon
- abstract@[open-review](#): Computer simulations have proven a valuable tool for understanding complex phenomena across the sciences. However, the utility of simulators for modelling and forecasting purposes is often restricted by low data quality, as well as practical limits to model fidelity. In order to circumvent these difficulties, we argue that modellers must treat simulators as idealistic representations of the true data generating process, and consequently should thoughtfully consider the risk of model misspecification. In this work we revisit neural posterior estimation (NPE), a class of algorithms that enable black-box parameter inference in simulation models, and consider the implication of a simulation-to-reality gap. While recent works have demonstrated reliable performance of these methods, the analyses have been performed using synthetic data generated by the simulator model itself, and have therefore only addressed the well-specified case. In this paper, we find that the presence of misspecification, in contrast, leads to unreliable inference when NPE is used naively. As a remedy we argue that principled scientific inquiry with simulators should incorporate a model criticism component, to facilitate interpretable identification of misspecification and a robust inference component, to fit "wrong but useful" models. We propose robust neural posterior estimation (RNPE), an extension of NPE to simultaneously achieve both these aims, through explicitly modelling the discrepancies between simulations and the observed data. We assess the approach on a range of artificially misspecified examples, and find RNPE performs well across the tasks, whereas naively using NPE leads to misleading and erratic posteriors.

## [Unsupervised Object Representation Learning using Translation and Rotation Group Equivariant VAE](#)

- Alireza Nasiri · Tristan Bepler
- abstract@[open-review](#): In many imaging modalities, objects of interest can occur in a variety of locations and poses (i.e. are subject to translations and rotations in 2d or 3d), but the location and pose of an object does not change its semantics (i.e. the object's essence). That is, the specific location and rotation of an airplane in satellite imagery, or the 3d rotation of a chair in a natural image, or the rotation of a particle in a cryo-electron micrograph, do not change the intrinsic nature of those objects. Here, we consider the problem of learning semantic representations of objects that are invariant to pose and location in a fully unsupervised manner. We address shortcomings in previous approaches to this problem by introducing TARGET-VAE, a translation and rotation group-equivariant variational autoencoder framework. TARGET-VAE combines three core innovations: 1) a rotation and translation group-equivariant encoder architecture, 2) a structurally disentangled distribution over latent rotation, translation, and a rotation-translation-invariant semantic object vector, which are jointly inferred by the approximate inference network, and 3) a spatially equivariant generator network. In comprehensive experiments, we show that TARGET-VAE learns disentangled representations without supervision that significantly improve upon and avoid the pathologies of previous methods. Semantic representations learned by TARGET-VAE on images highly corrupted by rotation and translation approach those learned on consistently posed objects, dramatically improving clustering and pose inference on multiple datasets.

## [Estimation of Entropy in Constant Space with Improved Sample Complexity](#)

- Maryam Aliakbarpour · Andrew McGregor · Jelani Nelson · Erik Waingarten
- abstract@[open-review](#): Recent work of Acharya et al.~(NeurIPS 2019) showed how to estimate the entropy of a distribution  $\\mathcal{D}$  over an alphabet of size  $k$  up to  $\\pm\\epsilon$  additive error by streaming over  $(k/\\epsilon)^3 \\cdot \\text{polylog}(1/\\epsilon)$  i.i.d. samples and using only  $O(1)$  words of memory. In this work, we give a new constant memory scheme that reduces the sample complexity to  $(k/\\epsilon)^2 \\cdot \\text{polylog}(1/\\epsilon)$ . We conjecture that this is optimal up to  $\\text{polylog}(1/\\epsilon)$  factors.

## [A Fast Scale-Invariant Algorithm for Non-negative Least Squares with Non-negative Data](#)

- Jelena Diakonikolas · Chenghui Li · Swati Padmanabhan · Chaobing Song
- abstract@[open-review](#): Nonnegative (linear) least square problems are a fundamental class of problems that is well-studied in statistical learning and for which solvers have been implemented in many of the standard programming languages used within the machine learning community. The existing off-the-

shelf solvers view the non-negativity constraint in these problems as an obstacle and, compared to unconstrained least squares, perform additional effort to address it. However, in many of the typical applications, the data itself is nonnegative as well, and we show that the nonnegativity in this case makes the problem easier. In particular, while the worst-case dimension-independent oracle complexity of unconstrained least squares problems necessarily scales with one of the data matrix constants (typically the spectral norm) and these problems are solved to additive error, we show that nonnegative least squares problems with nonnegative data are solvable to multiplicative error and with complexity that is independent of any matrix constants. The algorithm we introduce is accelerated and based on a primal-dual perspective. We further show how to provably obtain linear convergence using adaptive restart coupled with our method and demonstrate its effectiveness on large-scale data via numerical experiments.

## [A deep learning toolbox for stochastic stabilized supralinear networks](#)

- Wayne Soo · Mate Lengyel
- abstract@[open-review](#): There continues to be a trade-off between the biological realism and performance of neural networks. Contemporary deep learning techniques allow neural networks to be trained to perform challenging computations at (near) human-level, but these networks typically violate key biological constraints. More detailed models of biological neural networks can incorporate many of these constraints but typically suffer from subpar performance and trainability. Here, we narrow this gap by developing an effective method for training a canonical model of cortical neural circuits, the stabilized supralinear network (SSN), that in previous work had to be constructed manually or trained with undue constraints. SSNs are particularly challenging to train for the same reasons that make them biologically realistic: they are characterized by strongly-connected excitatory cells and expansive firing rate non-linearities that together make them prone to dynamical instabilities unless stabilized by appropriately tuned recurrent inhibition. Our method avoids such instabilities by initializing a small network and gradually increasing network size via the dynamics-neutral addition of neurons during training. We first show how SSNs can be trained to perform typical machine learning tasks by training an SSN on MNIST classification. We then demonstrate the effectiveness of our method by training an SSN on the challenging task of performing amortized Markov chain Monte Carlo-based inference under a Gaussian scale mixture generative model of natural image patches with a rich and diverse set of basis functions -- something that was not possible with previous methods. These results open the way to training realistic cortical-like neural networks on challenging tasks at scale.

## [Batch Bayesian Optimization on Permutations using the Acquisition Weighted Kernel](#)

- Changyong Oh · Roberto Bondesan · Efstratios Gavves · Max Welling
- abstract@[open-review](#): In this work we propose a batch Bayesian optimization method for combinatorial problems on permutations, which is well suited for expensive-to-evaluate objectives. We first introduce LAW, an efficient batch acquisition method based on determinantal point processes using the acquisition weighted kernel. Relying on multiple parallel evaluations, LAW enables accelerated search on combinatorial spaces. We then apply the framework to permutation problems, which have so far received little attention in the Bayesian Optimization literature, despite their practical importance. We call this method LAW2ORDER. On the theoretical front, we prove that LAW2ORDER has vanishing simple regret by showing that the batch cumulative regret is sublinear. Empirically, we assess the method on several standard combinatorial problems involving permutations such as quadratic assignment, flowshop scheduling and the traveling salesman, as well as on a structure learning task.

## [Aligning individual brains with fused unbalanced Gromov Wasserstein](#)

- Alexis Thual · Quang Huy TRAN · Tatiana Zemskova · Nicolas Courty · RÃ©mi Flamary · Stanislas Dehaene · Bertrand Thirion
- abstract@[open-review](#): Individual brains vary in both anatomy and functional organization, even within a given species. Inter-individual variability is a major impediment when trying to draw generalizable conclusions from neuroimaging data collected on groups of subjects. Current co-registration procedures rely on limited data, and thus lead to very coarse inter-subject alignments. In this work, we present a novel method for inter-subject alignment based on Optimal Transport, denoted as Fused Unbalanced Gromov Wasserstein (FUGW). The method aligns two cortical surfaces based on the similarity of their functional signatures in response to a variety of stimuli, while penalizing large deformations of individual topographic organization. We demonstrate that FUGW is suited for whole-brain landmark-free alignment. The unbalanced feature allows to deal with the fact that functional areas vary in size across subjects. Results show that FUGW alignment significantly increases between-subject correlation of activity during new independent fMRI tasks and runs, and leads to more precise maps of fMRI results at the group level.

## [VLMo: Unified Vision-Language Pre-Training with Mixture-of-Modality-Experts](#)

- Hangbo Bao · Wenhui Wang · Li Dong · Qiang Liu · Owais Khan Mohammed · Kriti Aggarwal · Songhao Piao · Subhrojit Som · Furu Wei
- abstract@[open-review](#): We present a unified Vision-Language pretrained Model (VLMo) that jointly learns a dual encoder and a fusion encoder with a modular Transformer network. Specifically, we introduce Mixture-of-Modality-Experts (MoME) Transformer, where each block contains a pool of modality-specific experts and a shared self-attention layer. Because of the modeling flexibility of MoME, pretrained VLMo can be fine-tuned as a fusion encoder for vision-language classification tasks, or used as a dual encoder for efficient image-text retrieval. Moreover, we propose a stagewise pre-training strategy, which effectively leverages large-scale image-only and text-only data besides image-text pairs. Experimental results show that VLMo achieves state-of-the-art results on various vision-language tasks, including VQA, NLVR2 and image-text retrieval.

## [Communication Efficient Federated Learning for Generalized Linear Bandits](#)

- Chuanhao Li · Hongning Wang
- abstract@[open-review](#): Contextual bandit algorithms have been recently studied under the federated learning setting to satisfy the demand of keeping data decentralized and pushing the learning of bandit models to the client side. But limited by the required communication efficiency, existing solutions are restricted to linear models to exploit their closed-form solutions for parameter estimation. Such a restricted model choice greatly hampers these algorithms' practical utility. In this paper, we take the first step to addressing this challenge by studying generalized linear bandit models under the federated learning setting. We propose a communication-efficient solution framework that employs online regression for local update and offline regression for global update. We rigorously proved, though the setting is more general and challenging, our algorithm can attain sub-linear rate in both regret and communication cost, which is also validated by our extensive empirical evaluations.

## [Grounding Aleatoric Uncertainty in Unsupervised Environment Design](#)

- Minqi Jiang · Michael Dennis · Jack Parker-Holder · Andrei Lupu · Heinrich KÃ¼ttler · Edward Grefenstette · Tim RocktÃ¤schel · Jakob Foerster
- abstract@[open-review](#): Adaptive curricula in reinforcement learning (RL) have proven effective for producing policies robust to discrepancies between the train and test environment. Recently, the Unsupervised Environment Design (UED) framework generalized RL curricula to generating sequences of entire environments, leading to new methods with robust minimax regret properties. Problematically, in partially-observable or stochastic settings, optimal policies may depend on the ground-truth distribution over aleatoric parameters of the environment in the intended deployment setting, while curriculum learning necessarily shifts the training distribution. We formalize this phenomenon as curriculum-induced covariate shift (CICS), and describe how its occurrence in aleatoric parameters can lead to suboptimal policies. Directly sampling these parameters from the ground-truth distribution avoids the issue, but thwarts curriculum learning. We propose SAMPLR, a minimax regret UED method that optimizes the ground-truth utility function, even when the underlying training data is biased due to CICS. We prove, and validate on challenging domains, that our approach preserves optimality under the ground-truth distribution, while promoting robustness across the full range of environment settings.

## Sequential Latent Variable Models for Multiagent Trajectories

- Dennis Fassmeyer · Pascal Fassmeyer · Ulf Brefeld
- abstract@[open-review](#): Analyzing the spatiotemporal behavior of multiple agents is of great interest to many communities. Existing probabilistic models in this regime employ either unsupervised generative settings, in which the latent space is described fully by discrete or continuous representations, or, are alternatively formalized in a (fully) supervised framework where weakly preserved labels add explicit information to a continuous latent representation learned from the data. To overcome the resulting limitations, we propose a novel objective function for processing multi-agent trajectories based on semi-supervised variational autoencoders, where equivariance and interaction of agents are modeled via customized graph networks. Our formulation disentangles discrete and continuous effects and allows discrete behavioral indicators in arbitrary quantity and annotation type to guide the generation process. This lifts applicability to relevant prediction problems beyond the generation of collective movements and provides an effective solution to incorporate expensive domain knowledge into interactive multi-agent systems. Empirically, our model outperforms various state-of-the-art baselines in generating future agent movements on interactive real-world datasets. We also show that our approach effectively learns to leverage unsupervised multi-agent sequences to improve classification of long-term locations as well as manually annotated situations on sports tracking data.

## Bayesian inference via sparse Hamiltonian flows

- Naitong Chen · Zuheng Xu · Trevor Campbell
- abstract@[open-review](#): A Bayesian coresnet is a small, weighted subset of data that replaces the full dataset during Bayesian inference, with the goal of reducing computational cost. Although past work has shown empirically that there often exists a coresnet with low inferential error, efficiently constructing such a coresnet remains a challenge. Current methods tend to be slow, require a secondary inference step after coresnet construction, and do not provide bounds on the data marginal evidence. In this work, we introduce a new method---sparse Hamiltonian flows---that addresses all three of these challenges. The method involves first subsampling the data uniformly, and then optimizing a Hamiltonian flow parametrized by coresnet weights and including periodic momentum quasi-refreshment steps. Theoretical results show that the method enables an exponential compression of the dataset in a representative model, and that the quasi-refreshment steps reduce the KL divergence to the target. Real and synthetic experiments demonstrate that sparse Hamiltonian flows provide accurate posterior approximations with significantly reduced runtime compared with competing dynamical-system-based inference methods.

## Learning with convolution and pooling operations in kernel methods

- Theodor Misiakiewicz · Song Mei
- abstract@[open-review](#): Recent empirical work has shown that hierarchical convolutional kernels inspired by convolutional neural networks (CNNs) significantly improve the performance of kernel methods in image classification tasks. A widely accepted explanation for their success is that these architectures encode hypothesis classes that are suitable for natural images. However, understanding the precise interplay between approximation and generalization in convolutional architectures remains a challenge. In this paper, we consider the stylized setting of covariates (image pixels) uniformly distributed on the hypercube, and characterize exactly the RKHS of kernels composed of single layers of convolution, pooling, and downsampling operations. We use this characterization to compute sharp asymptotics of the generalization error for any given function in high-dimension. In particular, we quantify the gain in sample complexity brought by enforcing locality with the convolution operation and approximate translation invariance with average pooling. Notably, these results provide a precise description of how convolution and pooling operations trade off approximation with generalization power in one layer convolutional kernels.

## Learning sparse features can lead to overfitting in neural networks

- Francesco Cagnetta · Matthieu Wyart · Leonardo Petrini · Eric Vanden-Eijnden
- abstract@[open-review](#): It is widely believed that the success of deep networks lies in their ability to learn a meaningful representation of the features of the data. Yet, understanding when and how this feature learning improves performance remains a challenge: for example, it is beneficial for modern architectures trained to classify images, whereas it is detrimental for fully-connected networks trained for the same task on the same data. Here we propose an explanation for this puzzle, by showing that feature learning can perform worse than lazy training (via random feature kernel or the NTK) as the former can lead to a sparser neural representation. Although sparsity is known to be essential for learning anisotropic data, it is detrimental when the target function is constant or smooth along certain directions of input space. We illustrate this phenomenon in two settings: (i) regression of Gaussian random functions on the  $d$ -dimensional unit sphere and (ii) classification of benchmark datasets of images. For (i), we compute the scaling of the generalization error with number of training points, and show that methods that do not learn features generalize better, even when the dimension of the input space is large. For (ii), we show empirically that learning features can indeed lead to sparse and thereby less smooth representations of the image predictors. This fact is plausibly responsible for deteriorating the performance, which is known to be correlated with smoothness along diffeomorphisms.

## Data Augmentation for Compositional Data: Advancing Predictive Models of the Microbiome

- Elliott Gordon-Rodriguez · Thomas Quinn · John Cunningham
- abstract@[open-review](#): Data augmentation plays a key role in modern machine learning pipelines. While numerous augmentation strategies have been studied in the context of computer vision and natural language processing, less is known for other data modalities. Our work extends the success of data augmentation to compositional data, i.e., simplex-valued data, which is of particular interest in microbiology, geochemistry, and other applications. Drawing on key principles from compositional data analysis, such as the Aitchison geometry of the simplex and subcompositions, we define novel augmentation strategies for this data modality. Incorporating our data augmentations into standard supervised learning pipelines results in consistent performance gains across a wide range of standard benchmark datasets. In particular, we set a new state-of-the-art for key disease prediction tasks including colorectal cancer, type 2 diabetes, and Crohn's disease. In addition, our data augmentations enable us to define a novel contrastive learning model, which improves on previous representation learning approaches for microbiome compositional data.

## On the Theoretical Properties of Noise Correlation in Stochastic Optimization

- Aurelien Lucchi · Frank Proske · Antonio Orvieto · Francis Bach · Hans Kersting
- abstract@[open-review](#): Studying the properties of stochastic noise to optimize complex non-convex functions has been an active area of research in the field of machine learning. Prior work~\citet{zhou2019pgd, wei2019noise} has shown that the noise of stochastic gradient descent improves optimization by overcoming undesirable obstacles in the landscape. Moreover, injecting artificial Gaussian noise has become a popular idea to quickly escape saddle points. Indeed, in the absence of reliable gradient information, the noise is used to explore the landscape, but it is unclear what type of noise is optimal in terms of exploration ability. In order to narrow this gap in our knowledge, we study a general type of continuous-time non-Markovian process, based on fractional Brownian motion, that allows for the increments of the process to be correlated. This generalizes processes based on Brownian motion, such as the Ornstein-Uhlenbeck process. We demonstrate how to discretize such processes which gives rise to the new algorithm ``fPGD''. This method is a generalization of the known algorithms PGD and Anti-PGD~\citet{orvieto2022anti}. We study the properties of fPGD both theoretically and empirically, demonstrating that it possesses exploration abilities that, in some cases, are favorable over PGD and Anti-PGD. These results open the field to novel ways to exploit noise for training machine learning models.

## Learning in Distributed Contextual Linear Bandits Without Sharing the Context

- Osama Hanna · Lin Yang · Christina Fragouli
- abstract@[open-review](#): Contextual linear bandits is a rich and theoretically important model that has many practical applications. Recently, this setup gained a lot of interest in applications over wireless where communication constraints can be a performance bottleneck, especially when the contexts come from a large  $d$ -dimensional space. In this paper, we consider the distributed contextual linear bandit learning problem, where the agents who observe the contexts and take actions are geographically separated from the learner who performs the learning while not seeing the contexts. We assume that contexts are generated from a distribution and propose a method that uses  $\approx 5d$  bits per context for the case of unknown context distribution and  $0$  bits per context if the context distribution is known, while achieving nearly the same regret bound as if the contexts were directly observable. The former bound improves upon existing bounds by a  $\log(T)$  factor, where  $T$  is the length of the horizon, while the latter achieves information theoretical tightness.

## [Learning low-dimensional generalizable natural features from retina using a U-net](#)

- Siwei Wang · Benjamin Hoshal · Elizabeth de Laittre · Thierry Mora · Michael Berry · Stephanie Palmer
- abstract@[open-review](#): Much of sensory neuroscience focuses on sensory features that are chosen by the experimenter because they are thought to be behaviorally relevant to the organism. However, it is not generally known what these features are in complex, natural scenes. This work focuses on using the retinal encoding of natural movies to determine the presumably behaviorally-relevant features that the brain represents. It is prohibitive to parameterize a natural movie and its respective retinal encoding fully. We use time within a natural movie as a proxy for the whole suite of features evolving across the scene. We then use a task-agnostic deep architecture, an encoder-decoder, to model the retinal encoding process and characterize its representation of "time in the natural scene" in a compressed latent space. In our end-to-end training, an encoder learns a compressed latent representation from a large population of salamander retinal ganglion cells responding to natural movies, while a decoder samples from this compressed latent space to generate the appropriate movie frame. By comparing latent representations of retinal activity from three movies, we find that the retina performs transfer learning to encode time: the precise, low-dimensional representation of time learned from one movie can be used to represent time in a different movie, with up to 17ms resolution. We then show that static textures and velocity features of a natural movie are synergistic. The retina simultaneously encodes both to establish a generalizable, low-dimensional representation of time in the natural scene.

## [Parameter-free Dynamic Graph Embedding for Link Prediction](#)

- Jiahao Liu · Dongsheng Li · Hansu Gu · Tun Lu · Peng Zhang · Ning Gu
- abstract@[open-review](#): Dynamic interaction graphs have been widely adopted to model the evolution of user-item interactions over time. There are two crucial factors when modelling user preferences for link prediction in dynamic interaction graphs: 1) collaborative relationship among users and 2) user personalized interaction patterns. Existing methods often implicitly consider these two factors together, which may lead to noisy user modelling when the two factors diverge. In addition, they usually require time-consuming parameter learning with back-propagation, which is prohibitive for real-time user preference modelling. To this end, this paper proposes FreeGEM, a parameter-free dynamic graph embedding method for link prediction. Firstly, to take advantage of the collaborative relationships, we propose an incremental graph embedding engine to obtain user/item embeddings, which is an Online-Monitor-Offline architecture consisting of an Online module to approximately embed users/items over time, a Monitor module to estimate the approximation error in real time and an Offline module to calibrate the user/item embeddings when the online approximation errors exceed a threshold. Meanwhile, we integrate attribute information into the model, which enables FreeGEM to better model users belonging to some under represented groups. Secondly, we design a personalized dynamic interaction pattern modeller, which combines dynamic time decay with attention mechanism to model user short-term interests. Experimental results on two link prediction tasks show that FreeGEM can outperform the state-of-the-art methods in accuracy while achieving over 36X improvement in efficiency. All code and datasets can be found in <https://github.com/FudanCISL/FreeGEM>.

## [Subgroup Robustness Grows On Trees: An Empirical Baseline Investigation](#)

- Josh Gardner · Zoran Popovic · Ludwig Schmidt
- abstract@[open-review](#): Many methods for fair, robust, or disparity-minimizing machine learning have been proposed across the machine learning community, but thorough empirical evaluation of their subgroup robustness is lacking. In this work, we address this gap in the context of tabular data, where sensitive subgroups are clearly-defined, real-world fairness problems abound, and prior works often fail to compare to state-of-the-art tree-based tabular data models. We conduct an empirical study of several previously-proposed methods for fair and robust learning alongside state-of-the-art tree-based methods for tabular data and other baseline methods. Over experiments with more than \$300,000 model configurations on eight datasets, we show that tree-based methods have surprisingly strong subgroup robustness properties even when compared to robustness- and fairness-enhancing methods. We show that three families of metrics -- those derived from accuracy, robust losses, and fairness metrics -- show strong empirical agreement within their family, but display little or no correlation across these families for non-tree models. As a result, the best tree-based models tend to show good performance over a range of metrics, while robust or group-fair models can show brittleness, with significant performance differences across different metrics for a fixed model. Our work suggests that tree-based ensemble models make an effective baseline for tabular data, and are a sensible default when subgroup robustness is desired.

## [Redistricting via Local Fairness](#)

- Pankaj Agarwal · Shao-Heng Ko · Kamesh Munagala · Erin Taylor
- abstract@[open-review](#): In this paper, we propose to use the concept of local fairness for auditing and ranking redistricting plans. Given a redistricting plan, a deviating group is a population-balanced contiguous region in which a majority of individuals are of the same interest and in the minority of their respective districts in the given redistricting plan; such a set of individuals have a justified complaint with how the redistricting plan was drawn. A redistricting plan with no deviating groups is called locally fair. We show that the problem of auditing a given plan for local fairness is NP-complete. We present an MCMC approach for auditing as well as ranking redistricting plans. We also present a dynamic programming based algorithm for the auditing problem that we use to demonstrate the efficacy of our MCMC approach. Using these tools, we test local fairness on real-world election data, showing that it is indeed possible to find plans that are almost or exactly locally fair. Further, we show that such plans can be generated while sacrificing very little in terms of compactness and existing fairness measures such as competitiveness of the districts or seat shares.

## [Causally motivated multi-shortcut identification and removal](#)

- Jiayun Zheng · Maggie Makar
- abstract@[open-review](#): For predictive models to provide reliable guidance in decision making processes, they are often required to be accurate and robust to distribution shift. Shortcut learning--where a model relies on spurious correlations or shortcuts to predict the target label--undermines the robustness property, leading to models with poor out-of-distribution accuracy despite good in-distribution performance. Existing work on shortcut learning either assumes that the set of possible shortcuts is known a priori or is discoverable using interpretability methods such as saliency maps. Instead, we propose a two step approach to (1) efficiently identify relevant shortcuts, and (2) leverage the identified shortcuts to build models that are robust to distribution shifts. Our approach relies on having access to a (possibly) high dimensional set of auxiliary labels at training time, some of which correspond to possible shortcuts. We show both theoretically and empirically that our approach is able to identify a small sufficient set of shortcuts leading to more efficient predictors in finite samples.

## [MetaTeacher: Coordinating Multi-Model Domain Adaptation for Medical Image Classification](#)

- Zhenbin Wang · Mao Ye · Xiatian Zhu · Liuhan Peng · Liang Tian · Yingying Zhu
- abstract@[open-review](#): In medical image analysis, we often need to build an image recognition system for a target scenario with the access to small labeled data and abundant unlabeled data, as well as multiple related models pretrained on different source scenarios. This presents the combined challenges of multi-source-free domain adaptation and semi-supervised learning simultaneously. However, both problems are typically studied independently in the literature, and how to effectively combine existing methods is non-trivial in design. In this work, we introduce a novel MetaTeacher framework with three key components: (1) A learnable coordinating scheme for adaptive domain adaptation of individual source models, (2) A mutual feedback mechanism between the target model and source models for more coherent learning, and (3) A semi-supervised bilevel optimization algorithm for consistently organizing the adaption of source models and the learning of target model. It aims to leverage the knowledge of source models adaptively whilst maximize their complementary benefits collectively to counter the challenge of limited supervision. Extensive experiments on five chest x-ray image datasets show that our method outperforms clearly all the state-of-the-art alternatives.

## [Characterizing the Ventral Visual Stream with Response-Optimized Neural Encoding Models](#)

- Meenakshi Khosla · Keith Jamison · Amy Kuceyeski · Mert Sabuncu
- abstract@[open-review](#): Decades of experimental research based on simple, abstract stimuli has revealed the coding principles of the ventral visual processing hierarchy, from the presence of edge detectors in the primary visual cortex to the selectivity for complex visual categories in the anterior ventral stream. However, these studies are, by construction, constrained by their  $\{\text{a priori}\}$  hypotheses. Furthermore, beyond the early stages, precise neuronal tuning properties and representational transformations along the ventral visual pathway remain poorly understood. In this work, we propose to employ response-optimized encoding models trained solely to predict the functional MRI activation, in order to gain insights into the tuning properties and representational transformations in the series of areas along the ventral visual pathway. We demonstrate the strong generalization abilities of these models on artificial stimuli and novel datasets. Intriguingly, we find that response-optimized models trained towards the ventral-occipital and lateral-occipital areas, but not early visual areas, can recapitulate complex visual behaviors like object categorization and perceived image-similarity in humans. We further probe the trained networks to reveal representational biases in different visual areas and generate experimentally testable hypotheses. Our analyses suggest a shape-based processing along the ventral visual stream and provide a unified picture of multiple neural phenomena characterized over the last decades with controlled fMRI studies.

## [Learning Bipartite Graphs: Heavy Tails and Multiple Components](#)

- JosÃ© VinÃ¢cius de Miranda Cardoso · Jiaxi Ying · Daniel Palomar
- abstract@[open-review](#): We investigate the problem of learning an undirected, weighted bipartite graph under the Gaussian Markov random field model, for which we present an optimization formulation along with an efficient algorithm based on the projected gradient descent. Motivated by practical applications, where outliers or heavy-tailed events are present, we extend the proposed learning scheme to the case in which the data follow a multivariate Student-\$\\$ distribution. As a result, the optimization program is no longer convex, but a verifiably convergent iterative algorithm is proposed based on the majorization-minimization framework. Finally, we propose an efficient and provably convergent algorithm for learning \$k\$-component bipartite graphs that leverages rank constraints of the underlying graph Laplacian matrix. The proposed estimators outperform state-of-the-art methods for bipartite graph learning, as evidenced by real-world experiments using financial time series data.

## [Autoregressive Perturbations for Data Poisoning](#)

- Pedro Sandoval-Segura · Vasu Singla · Jonas Geiping · Micah Goldblum · Tom Goldstein · David Jacobs
- abstract@[open-review](#): The prevalence of data scraping from social media as a means to obtain datasets has led to growing concerns regarding unauthorized use of data. Data poisoning attacks have been proposed as a bulwark against scraping, as they make data ``unlearnable'' by adding small, imperceptible perturbations. Unfortunately, existing methods require knowledge of both the target architecture and the complete dataset so that a surrogate network can be trained, the parameters of which are used to generate the attack. In this work, we introduce autoregressive (AR) poisoning, a method that can generate poisoned data without access to the broader dataset. The proposed AR perturbations are generic, can be applied across different datasets, and can poison different architectures. Compared to existing unlearnable methods, our AR poisons are more resistant against common defenses such as adversarial training and strong data augmentations. Our analysis further provides insight into what makes an effective data poison.

## [Towards Trustworthy Automatic Diagnosis Systems by Emulating Doctors' Reasoning with Deep Reinforcement Learning](#)

- Arsene Fansi Tchango · Zhi Wen · Gaetan Marceau Caron · Joumana Ghosn · Rishab Goel · Julien Martel
- abstract@[open-review](#): To reduce medical doctor's workload and democratize access to medical care, the automation of the evidence acquisition and diagnosis process has attracted increasing attention recently. However, most works proposed in the machine learning literature focus solely on improving the prediction accuracy of the patient's pathology. We argue that this objective is insufficient to ensure doctors' acceptability of the system. For doctors to trust the system recommendations, they need to understand how the gathered evidences led to the predicted diseases. In particular, interactions between the system and a patient should emulate doctors' reasoning. To do so, we propose to model the evidence acquisition and automatic diagnosis tasks in a deep reinforcement learning framework by considering three essential aspects of doctors' reasoning, namely using differential diagnosis with the exploration-confirmation approach while prioritizing severe pathologies. We propose metrics for evaluating the interaction quality concerning these three aspects. We show that our approach performs better than existing models while maintaining competitive prediction accuracy.

## [Does GNN Pretraining Help Molecular Representation?](#)

- Ruoxi Sun · Hanjun Dai · Adams Yu
- abstract@[open-review](#): Extracting informative representations of molecules using Graph neural networks (GNNs) is crucial in AI-driven drug design and discovery. Recently, the graph research community has been trying to replicate the success of self-supervised pretraining in natural language processing, with several successes claimed. However, we find the benefit brought by self-supervised pretraining on molecular data can be negligible in many cases. We conduct thorough ablation studies on the key components of GNN pretraining, including pretraining objectives, data splitting methods, input features, pretraining dataset scales, and GNN architectures, in deciding the accuracy of the downstream tasks. Our first important finding is, self-supervised graph pretraining do not have statistically significant advantages over non-pretraining methods in many settings. Secondly, although improvement can be observed with additional supervised pretraining, the improvement may diminish with richer features or more balanced data splits. Thirdly, experimental hyper-parameters may have a larger impact on accuracy of downstream tasks than the choice of pretraining tasks. We hypothesize the complexity of pretraining on molecules is insufficient, leading to less transferable knowledge for downstream tasks.

## [Discrete Compositional Representations as an Abstraction for Goal Conditioned Reinforcement Learning](#)

- Riashat Islam · Hongyu Zang · Anirudh Goyal · Alex Lamb · Kenji Kawaguchi · Xin Li · Romain Laroche · Yoshua Bengio · Remi Tachet des Combes
- abstract@[open-review](#): Goal-conditioned reinforcement learning (RL) is a promising direction for training agents that are capable of solving multiple tasks and reach a diverse set of objectives. How to \textit{specify} and \textit{ground} these goals in such a way that we can both reliably reach goals during training as well as generalize to new goals during evaluation remains an open area of research. Defining goals in the space of noisy, high-dimensional sensory inputs is one possibility, yet this poses a challenge for training goal-conditioned agents, or even for generalization to novel goals. We propose to address this by learning compositional representations of goals and processing the resulting representation via a discretization bottleneck, for

coarser specification of goals, through an approach we call DGRL. We show that discretizing outputs from goal encoders through a bottleneck can work well in goal-conditioned RL setups, by experimentally evaluating this method on tasks ranging from maze environments to complex robotic navigation and manipulation tasks. Additionally, we show a theoretical result which bounds the expected return for goals not observed during training, while still allowing for specifying goals with expressive combinatorial structure.

## [SKFlow: Learning Optical Flow with Super Kernel Sizes](#)

- SHANGKUN SUN · Yuanqi Chen · Ge Li · Yu Zhu · Guodong Guo
- abstract@[open-review](#): Optical flow estimation is a classical yet challenging task in computer vision. One of the essential factors to accurately predict optical flow is to alleviate occlusions between frames. However, it is still a thorny problem for current top-performing optical flow estimation methods due to insufficient local evidence to model occluded areas. In this paper, we propose Super Kernel Flow Network (SKFlow), a CNN architecture to ameliorate the impacts of occlusions on optical flow estimation. SKFlow benefits from the super kernels which bring enlarged receptive fields to complement the absent matching information and recover the occluded motions. We present efficient super kernel designs by utilizing conical connections and hybrid depth-wise convolutions. Extensive experiments demonstrate the effectiveness of SKFlow on multiple benchmarks, especially on the occluded areas. Without pre-trained backbones on ImageNet and with modest increase in computation, SKFlow achieves compelling performance and ranks \$textbf{1st}\$ among current published methods on Sintel benchmark. On the challenging Sintel final pass test set, SKFlow attains the average end-point error of \$2.23\$, which surpasses the best published result \$2.47\$ by \$9.72\%\$.

## [E-MAPP: Efficient Multi-Agent Reinforcement Learning with Parallel Program Guidance](#)

- Can Chang · Ni Mu · Jiajun Wu · Ling Pan · Huazhe Xu
- abstract@[open-review](#): A critical challenge in multi-agent reinforcement learning(MARL) is for multiple agents to efficiently accomplish complex, long-horizon tasks. The agents often have difficulties in cooperating on common goals, dividing complex tasks, and planning through several stages to make progress. We propose to address these challenges by guiding agents with programs designed for parallelization, since programs as a representation contain rich structural and semantic information, and are widely used as abstractions for long-horizon tasks. Specifically, we introduce Efficient Multi-Agent Reinforcement Learning with Parallel Program Guidance(E-MAPP), a novel framework that leverages parallel programs to guide multiple agents to efficiently accomplish goals that require planning over \$10+\$ stages. E-MAPP integrates the structural information from a parallel program, promotes the cooperative behaviors grounded in program semantics, and improves the time efficiency via a task allocator. We conduct extensive experiments on a series of challenging, long-horizon cooperative tasks in the Overcooked environment. Results show that E-MAPP outperforms strong baselines in terms of the completion rate, time efficiency, and zero-shot generalization ability by a large margin.

## [Unsupervised Learning of Shape Programs with Repeatable Implicit Parts](#)

- Boyang Deng · Sumith Kulal · Zhengyang Dong · Congyue Deng · Yonglong Tian · Jiajun Wu
- abstract@[open-review](#): Shape programs encode shape structures by representing object parts as subroutines and constructing the overall shape by composing these subroutines. This usually involves the reuse of subroutines for repeatable parts, enabling the modeling of correlations among shape elements such as geometric similarity. However, existing learning-based shape programs suffer from limited representation capacity because they use coarse geometry representations such as geometric primitives and low-resolution voxel grids. Further, their training requires manually annotated ground-truth programs, which are expensive to attain. We address these limitations by proposing Shape Programs with Repeatable Implicit Parts (ProGRIP). Using implicit functions to represent parts, ProGRIP greatly boosts the representation capacity of shape programs while preserving the higher-level structure of repetitions and symmetry. Meanwhile, we free ProGRIP from any inaccessible supervised training via devising a matching-based unsupervised training objective. Our empirical studies show that ProGRIP outperforms existing structured representations in shape reconstruction fidelity as well as segmentation accuracy of semantic parts.

## [A Fourier Approach to Mixture Learning](#)

- Mingda Qiao · Guru Guruganesh · Ankit Rawat · Kumar Avinava Dubey · Manzil Zaheer
- abstract@[open-review](#): We revisit the problem of learning mixtures of spherical Gaussians. Given samples from a mixture \$\frac{1}{k} \sum\_{j=1}^k \mathcal{N}(\mu\_j, I\_d)\$, the goal is to estimate the means \$\mu\_1, \mu\_2, \dots, \mu\_k \in \mathbb{R}^d\$ up to a small error. The hardness of this learning problem can be measured by the separation \$\Delta\$ defined as the minimum distance between all pairs of means. Regev and Vijayaraghavan (2017) showed that with \$\Delta = \Omega(\sqrt{\log k})\$ separation, the means can be learned using \$\mathrm{poly}(k, d)\$ samples, whereas super-polynomially many samples are required if \$\Delta = o(\sqrt{\log k})\$ and \$d = \Omega(\log k)\$. This leaves open the low-dimensional regime where \$d = o(\log k)\$. In this work, we give an algorithm that efficiently learns the means in \$d = O(\log k / \log \log k)\$ dimensions under separation \$d / \sqrt{\log k}\$ (modulo doubly logarithmic factors). This separation is strictly smaller than \$\sqrt{\log k}\$, and is also shown to be necessary. Along with the results of Regev and Vijayaraghavan (2017), our work almost pins down the critical separation threshold at which efficient parameter learning becomes possible for spherical Gaussian mixtures. More generally, our algorithm runs in time \$\mathrm{poly}(k) \cdot f(d, \Delta, \epsilon)\$, and is thus fixed-parameter tractable in parameters \$d\$, \$\Delta\$ and \$\epsilon\$. Our approach is based on estimating the Fourier transform of the mixture at carefully chosen frequencies, and both the algorithm and its analysis are simple and elementary. Our positive results can be easily extended to learning mixtures of non-Gaussian distributions, under a mild condition on the Fourier spectrum of the distribution.

## [Look Around and Refer: 2D Synthetic Semantics Knowledge Distillation for 3D Visual Grounding](#)

- eslam mohamed · Yasmeen Alsaedy · Mohamed Elhoseiny
- abstract@[open-review](#): 3D visual grounding task has been explored with visual and language streams to comprehend referential language for identifying targeted objects in 3D scenes. However, most existing methods devote the visual stream to capture the 3D visual clues using off-the-shelf point clouds encoders. The main question we address is ‘can we consolidate the 3D visual stream by 2D clues and efficiently utilize them in both training and testing phases?’. The main idea is to assist the 3D encoder by incorporating rich 2D object representations without requiring extra 2D inputs. To this end, we leverage 2D clues, synthetically generated from 3D point clouds, that empirically show their aptitude to boost the quality of the learned visual representations. We validate our approach through comprehensive experiments on Nr3D, Sr3D, and ScanRefer datasets. Our experiments show consistent performance gains against counterparts, where our proposed module, dubbed as LAR, significantly outperforms state-of-the-art 3D visual grounding techniques on three benchmarks. Our code will be made publicly available.

## [Transfer Learning on Heterogeneous Feature Spaces for Treatment Effects Estimation](#)

- Ioana Bica · Mihaela van der Schaar
- abstract@[open-review](#): Consider the problem of improving the estimation of conditional average treatment effects (CATE) for a target domain of interest by leveraging related information from a source domain with a different feature space. This heterogeneous transfer learning problem for CATE estimation is ubiquitous in areas such as healthcare where we may wish to evaluate the effectiveness of a treatment for a new patient population for which different clinical covariates and limited data are available. In this paper, we address this problem by introducing several building blocks that use representation learning to handle the heterogeneous feature spaces and a flexible multi-task architecture with shared and private layers to transfer information between potential outcome functions across domains. Then, we show how these building blocks can be used to recover transfer learning equivalents of the standard CATE learners. On a new semi-synthetic data simulation benchmark for heterogeneous transfer learning, we not only demonstrate performance

improvements of our heterogeneous transfer causal effect learners across datasets, but also provide insights into the differences between these learners from a transfer perspective.

## [The Minority Matters: A Diversity-Promoting Collaborative Metric Learning Algorithm](#)

- Shilong Bao · Qianqian Xu · Zhiyong Yang · Yuan He · Xiaochun Cao · Qingming Huang
- abstract@[open-review](#): Collaborative Metric Learning (CML) has recently emerged as a popular method in recommendation systems (RS), closing the gap between metric learning and Collaborative Filtering. Following the convention of RS, existing methods exploit unique user representation in their model design. This paper focuses on a challenging scenario where a user has multiple categories of interests. Under this setting, we argue that the unique user representation might induce preference bias, especially when the item category distribution is imbalanced. To address this issue, we propose a novel method called Diversity-Promoting Collaborative Metric Learning (DPCML), with the hope of considering the commonly ignored minority interest of the user. The key idea behind DPCML is to include a multiple set of representations for each user in the system. Based on this embedding paradigm, user preference toward an item is aggregated from different embeddings by taking the minimum item-user distance among the user embedding set. Furthermore, we observe that the diversity of the embeddings for the same user also plays an essential role in the model. To this end, we propose a diversity control regularization term to accommodate the multi-vector representation strategy better. Theoretically, we show that DPCML could generalize well to unseen test data by tackling the challenge of the annoying operation that comes from the minimum value. Experiments over a range of benchmark datasets speak to the efficacy of DPCML.

## [Generalization Error Bounds on Deep Learning with Markov Datasets](#)

- Lan V. Truong
- abstract@[open-review](#): In this paper, we derive upper bounds on generalization errors for deep neural networks with Markov datasets. These bounds are developed based on Koltchinskii and Panchenko's approach for bounding the generalization error of combined classifiers with i.i.d. datasets. The development of new symmetrization inequalities in high-dimensional probability for Markov chains is a key element in our extension, where the spectral gap of the infinitesimal generator of the Markov chain plays a key parameter in these inequalities. We also propose a simple method to convert these bounds and other similar ones in traditional deep learning and machine learning to Bayesian counterparts for both i.i.d. and Markov datasets. Extensions to \$m\$-order homogeneous Markov chains such as AR and ARMA models and mixtures of several Markov data services are given.

## [Online Agnostic Multiclass Boosting](#)

- Vinod Raman · Ambuj Tewari
- abstract@[open-review](#): Boosting is a fundamental approach in machine learning that enjoys both strong theoretical and practical guarantees. At a high-level, boosting algorithms cleverly aggregate weak learners to generate predictions with arbitrarily high accuracy. In this way, boosting algorithms convert weak learners into strong ones. Recently, Brukhim et al. [2020] extended boosting to the online agnostic binary classification setting. A key ingredient in their approach is a clean and simple reduction to online convex optimization, one that efficiently converts an arbitrary online convex optimizer to an agnostic online booster. In this work, we extend this reduction to multiclass problems and give the first boosting algorithm for online agnostic multiclass classification. Our reduction also enables the construction of algorithms for statistical agnostic, online realizable, and statistical realizable multiclass boosting.

## [BOME! Bilevel Optimization Made Easy: A Simple First-Order Approach](#)

- Mao Ye · Bo Liu · Stephen Wright · Peter Stone · Qiang Liu
- abstract@[open-review](#): Bilevel optimization (BO) is useful for solving a variety of important machine learning problems including but not limited to hyperparameter optimization, meta-learning, continual learning, and reinforcement learning. Conventional BO methods need to differentiate through the low-level optimization process with implicit differentiation, which requires expensive calculations related to the Hessian matrix. There has been a recent quest for first-order methods for BO, but the methods proposed to date tend to be complicated and impractical for large-scale deep learning applications. In this work, we propose a simple first-order BO algorithm that depends only on first-order gradient information, requires no implicit differentiation, and is practical and efficient for large-scale non-convex functions in deep learning. We provide non-asymptotic convergence analysis of the proposed method to stationary points for non-convex objectives and present empirical results that show its superior practical performance.

## [Cluster Randomized Designs for One-Sided Bipartite Experiments](#)

- Jennifer Brennan · Vahab Mirrokni · Jean Pouget-Abadie
- abstract@[open-review](#): The conclusions of randomized controlled trials may be biased when the outcome of one unit depends on the treatment status of other units, a problem known as \textit{interference}. In this work, we study interference in the setting of one-sided bipartite experiments in which the experimental units---where treatments are randomized and outcomes are measured---do not interact directly. Instead, their interactions are mediated through their connections to \textit{interference units} on the other side of the graph. Examples of this type of interference are common in marketplaces and two-sided platforms. The \textit{cluster-randomized design} is a popular method to mitigate interference when the graph is known, but it has not been well-studied in the one-sided bipartite experiment setting. In this work, we formalize a natural model for interference in one-sided bipartite experiments using the exposure mapping framework. We first exhibit settings under which existing cluster-randomized designs fail to properly mitigate interference under this model. We then show that minimizing the bias of the difference-in-means estimator under our model results in a balanced partitioning clustering objective with a natural interpretation. We further prove that our design is minimax optimal over the class of linear potential outcomes models with bounded interference. We conclude by providing theoretical and experimental evidence of the robustness of our design to a variety of interference graphs and potential outcomes models.

## [Evaluating Robustness to Dataset Shift via Parametric Robustness Sets](#)

- Michael Oberst · Nikolaj Thams · David Sontag
- abstract@[open-review](#): We give a method for proactively identifying small, plausible shifts in distribution which lead to large differences in model performance. These shifts are defined via parametric changes in the causal mechanisms of observed variables, where constraints on parameters yield a "robustness set" of plausible distributions and a corresponding worst-case loss over the set. While the loss under an individual parametric shift can be estimated via reweighting techniques such as importance sampling, the resulting worst-case optimization problem is non-convex, and the estimate may suffer from large variance. For small shifts, however, we can construct a local second-order approximation to the loss under shift and cast the problem of finding a worst-case shift as a particular non-convex quadratic optimization problem, for which efficient algorithms are available. We demonstrate that this second-order approximation can be estimated directly for shifts in conditional exponential family models, and we bound the approximation error. We apply our approach to a computer vision task (classifying gender from images), revealing sensitivity to shifts in non-causal attributes.

## [Masked Prediction: A Parameter Identifiability View](#)

- Bingbin Liu · Daniel Hsu · Pradeep Ravikumar · Andrej Risteski

- abstract@[open-review](#): The vast majority of work in self-supervised learning have focused on assessing recovered features by a chosen set of downstream tasks. While there are several commonly used benchmark datasets, this lens of feature learning requires assumptions on the downstream tasks which are not inherent to the data distribution itself. In this paper, we present an alternative lens, one of parameter identifiability: assuming data comes from a parametric probabilistic model, we train a self-supervised learning predictor with a suitable parametric form, and ask whether the parameters of the optimal predictor can be used to extract the parameters of the ground truth generative model. Specifically, we focus on latent-variable models capturing sequential structures, namely Hidden Markov Models with both discrete and conditionally Gaussian observations. We focus on masked prediction as the self-supervised learning task and study the optimal masked predictor. We show that parameter identifiability is governed by the task difficulty, which is determined by the choice of data model and the amount of tokens to predict. Technique-wise, we uncover close connections with the uniqueness of tensor rank decompositions, a widely used tool in studying identifiability through the lens of moments.

## [\*\*When does return-conditioned supervised learning work for offline reinforcement learning?\*\*](#)

- David Brandfonbrener Â· Alberto Bietti Â· Jacob Buckman Â· Romain Laroche Â· Joan Bruna
- abstract@[open-review](#): Several recent works have proposed a class of algorithms for the offline reinforcement learning (RL) problem that we will refer to as return-conditioned supervised learning (RCSL). RCSL algorithms learn the distribution of actions conditioned on both the state and the return of the trajectory. Then they define a policy by conditioning on achieving high return. In this paper, we provide a rigorous study of the capabilities and limitations of RCSL something which is crucially missing in previous work. We find that RCSL returns the optimal policy under a set of assumptions that are stronger than those needed for the more traditional dynamic programming-based algorithms. We provide specific examples of MDPs and datasets that illustrate the necessity of these assumptions and the limits of RCSL. Finally, we present empirical evidence that these limitations will also cause issues in practice by providing illustrative experiments in simple point-mass environments and on datasets from the D4RL benchmark.

## [\*\*PAC Prediction Sets for Meta-Learning\*\*](#)

- Sangdon Park Â· Edgar Dobriban Â· Insup Lee Â· Osbert Bastani
- abstract@[open-review](#): Uncertainty quantification is a key component of machine learning models targeted at safety-critical systems such as in healthcare or autonomous vehicles. We study this problem in the context of meta learning, where the goal is to quickly adapt a predictor to new tasks. In particular, we propose a novel algorithm to construct \emph{PAC prediction sets}, which capture uncertainty via sets of labels, that can be adapted to new tasks with only a few training examples. These prediction sets satisfy an extension of the typical PAC guarantee to the meta learning setting; in particular, the PAC guarantee holds with high probability over future tasks. We demonstrate the efficacy of our approach on four datasets across three application domains: mini-ImageNet and CIFAR10-C in the visual domain, FewRel in the language domain, and the CDC Heart Dataset in the medical domain. In particular, our prediction sets satisfy the PAC guarantee while having smaller size compared to other baselines that also satisfy this guarantee.

## [\*\*Policy Optimization with Advantage Regularization for Long-Term Fairness in Decision Systems\*\*](#)

- Eric Yu Â· Zhizhen Qin Â· Min Kyung Lee Â· Sicun Gao
- abstract@[open-review](#): Long-term fairness is an important factor of consideration for designing and deploying learning-based decision systems that do not discriminate against certain groups or populations in high-impact decision-making contexts, such as for admission and hiring, criminal justice, and resource allocation. Recent work has used the framework of Markov Decision Processes (MDPs) to formulate decision-making with long-term fairness requirements in dynamically changing environments, and demonstrated major challenges in directly deploying heuristic and rule-based policies that worked well in static environments. We will show that policy optimization methods from deep reinforcement learning can be used to find strictly better decision policies represented by neural networks, which can often achieve both higher overall utility and less violation of the fairness requirements, compared to previously known strategies. In particular, we propose a new method for imposing fairness requirements in policy optimization algorithms by regularizing the advantage evaluation in policy gradient, which makes it easy to impose fairness constraints without reward engineering or sacrificing training efficiency. We perform a detailed evaluation of the proposed methods in three established case studies, including attention allocation in incident monitoring, bank loan approval, and vaccine distribution for infectious diseases in population networks.

## [\*\*Layer Freezing & Data Sieving: Missing Pieces of a Generic Framework for Sparse Training\*\*](#)

- Geng Yuan Â· Yanyu Li Â· Sheng Li Â· Zhenglun Kong Â· Sergey Tulyakov Â· Xulong Tang Â· Yanzhi Wang Â· Jian Ren
- abstract@[open-review](#): Recently, sparse training has emerged as a promising paradigm for efficient deep learning on edge devices. The current research mainly devotes the efforts to reducing training costs by further increasing model sparsity. However, increasing sparsity is not always ideal since it will inevitably introduce severe accuracy degradation at an extremely high sparsity level. This paper intends to explore other possible directions to effectively and efficiently reduce sparse training costs while preserving accuracy. To this end, we investigate two techniques, namely, layer freezing and data sieving. First, the layer freezing approach has shown its success in dense model training and fine-tuning, yet it has never been adopted in the sparse training domain. Nevertheless, the unique characteristics of sparse training may hinder the incorporation of layer freezing techniques. Therefore, we analyze the feasibility and potentiality of using the layer freezing technique in sparse training and find it has the potential to save considerable training costs. Second, we propose a data sieving method for dataset-efficient training, which further reduces training costs by ensuring only a partial dataset is used throughout the entire training process. We show that both techniques can be well incorporated into the sparse training algorithm to form a generic framework, which we dub SpFDE. Our extensive experiments demonstrate that SpFDE can significantly reduce training costs while preserving accuracy from three dimensions: weight sparsity, layer freezing, and dataset sieving. Our code and models will be released.

## [\*\*Modular Flows: Differential Molecular Generation\*\*](#)

- Yogesh Verma Â· Samuel Kaski Â· Markus Heinonen Â· Vikas Garg
- abstract@[open-review](#): Generating new molecules is fundamental to advancing critical applications such as drug discovery and material synthesis. Flows can generate molecules effectively by inverting the encoding process, however, existing flow models either require artifactual dequantization or specific node/edge orderings, lack desiderata such as permutation invariance, or induce discrepancy between encoding and decoding steps that necessitates post hoc validity correction. Inspired by graph PDEs, we circumvent these issues with novel continuous normalizing E(3)-equivariant flows, based on a system of coupled node ODEs, that repeatedly reconcile locally toward globally aligned densities. Our models can be cast as message passing temporal networks, and result in superlative density estimation and molecular generation. In particular, our generated samples achieve state of the art on both the standard QM9 and ZINC250K benchmarks.

## [\*\*Compositional generalization through abstract representations in human and artificial neural networks\*\*](#)

- Takuya Ito Â· Tim Klinger Â· Doug Schultz Â· John Murray Â· Michael Cole Â· Mattia Rigotti
- abstract@[open-review](#): Humans have a remarkable ability to rapidly generalize to new tasks that is difficult to reproduce in artificial learning systems. Compositional generalization has been proposed as a key mechanism supporting generalization in humans, but evidence of its neural implementation and impact on behavior is still scarce. Here we study the computational properties associated with compositional generalization in both humans and artificial neural networks (ANNs) on a highly compositional task. First, we identified behavioral signatures of compositional generalization in humans, along with their neural correlates using whole-cortex functional magnetic resonance imaging (fMRI) data. Next, we designed pretraining paradigms aided by a procedure we term primitives pretraining to endow compositional task elements into ANNs. We found that ANNs with this prior knowledge had greater correspondence with human behavior and neural compositional signatures. Importantly, primitives pretraining induced abstract internal representations,

excellent zero-shot generalization, and sample-efficient learning. Moreover, it gave rise to a hierarchy of abstract representations that matched human fMRI data, where sensory rule abstractions emerged in early sensory areas, and motor rule abstractions emerged in later motor areas. Our findings give empirical support to the role of compositional generalization in humans behavior, implicate abstract representations as its neural implementation, and illustrate that these representations can be embedded into ANNs by designing simple and efficient pretraining procedures.

## [Large \(robust\) models from computational constraints](#)

- Sanjam Garg · Somesh Jha · Saeed Mahloujifar · Mohammad Mahmoody · Mingyuan Wang
- abstract@[open-review](#): Large models with thousands of parameters have been hugely successful. In this work, we ask: can the need for large models be due to the computational limitation of the learner? Additionally, we ask, is this situation exacerbated for robust learning? We show that this indeed could be the case. We show learning tasks for which computationally bounded learners need significantly more model parameters than those of information-theoretic learners. Furthermore, we show that even more model parameters could be necessary for the case of robust learning. In particular, for computationally bounded learners, we boost the recent result of Bubeck and Sellke [NeurIPS'2021], which shows that robust models might need more parameters, to the computational regime and show that bounded learners could provably need an even larger number of parameters. Next, we ask: can we hope to remedy the situation for robust computationally bounded learning by restricting adversaries to also be computationally bounded? Here again, we show that this might be possible. Specifically, building on Garg, Jha, Mahloujifar, and Mahmoody [ALT'2020], we demonstrate a learning task that can be learned robustly and efficiently against a computationally bounded attacker with a small model. On the other hand, to be robust against an information-theoretic attacker requires the learner to output a much larger model.

## [Change-point Detection for Sparse and Dense Functional Data in General Dimensions](#)

- Carlos Misael Madrid Padilla · Daren Wang · Zifeng Zhao · Yi Yu
- abstract@[open-review](#): We study the problem of change-point detection and localisation for functional data sequentially observed on a general  $d$ -dimensional space, where we allow the functional curves to be either sparsely or densely sampled. Data of this form naturally arise in a wide range of applications such as biology, neuroscience, climatology and finance. To achieve such a task, we propose a kernel-based algorithm named functional seeded binary segmentation (FSBS). FSBS is computationally efficient, can handle discretely observed functional data, and is theoretically sound for heavy-tailed and temporally-dependent observations. Moreover, FSBS works for a general  $d$ -dimensional domain, which is the first in the literature of change-point estimation for functional data. We show the consistency of FSBS for multiple change-point estimation and further provide a sharp localisation error rate, which reveals an interesting phase transition phenomenon depending on the number of functional curves observed and the sampling frequency for each curve. Extensive numerical experiments illustrate the effectiveness of FSBS and its advantage over existing methods in the literature under various settings. A real data application is further conducted, where FSBS localises change-points of sea surface temperature patterns in the south Pacific attributed to El Niño.

## [Mesoscopic modeling of hidden spiking neurons](#)

- Shuqi Wang · Valentin Schmutz · Guillaume Bellec · Wulfram Gerstner
- abstract@[open-review](#): Can we use spiking neural networks (SNN) as generative models of multi-neuronal recordings, while taking into account that most neurons are unobserved? Modeling the unobserved neurons with large pools of hidden spiking neurons leads to severely underconstrained problems that are hard to tackle with maximum likelihood estimation. In this work, we use coarse-graining and mean-field approximations to derive a bottom-up, neuronally-grounded latent variable model (neuLVM), where the activity of the unobserved neurons is reduced to a low-dimensional mesoscopic description. In contrast to previous latent variable models, neuLVM can be explicitly mapped to a recurrent, multi-population SNN, giving it a transparent biological interpretation. We show, on synthetic spike trains, that a few observed neurons are sufficient for neuLVM to perform efficient model inversion of large SNNs, in the sense that it can recover connectivity parameters, infer single-trial latent population activity, reproduce ongoing metastable dynamics, and generalize when subjected to perturbations mimicking photo-stimulation.

## [A Near-Optimal Best-of-Both-Worlds Algorithm for Online Learning with Feedback Graphs](#)

- Chloé Rouyer · Dirk van der Hoeven · Nicolò Cesa-Bianchi · Yevgeny Seldin
- abstract@[open-review](#): We consider online learning with feedback graphs, a sequential decision-making framework where the learner's feedback is determined by a directed graph over the action set. We present a computationally-efficient algorithm for learning in this framework that simultaneously achieves near-optimal regret bounds in both stochastic and adversarial environments. The bound against oblivious adversaries is  $\tilde{O}(\sqrt{\alpha T})$ , where  $T$  is the time horizon and  $\alpha$  is the independence number of the feedback graph. The bound against stochastic environments is  $O(\max(S \ln \mathcal{I}(G)) \sum_i \Delta_i)$  where  $\mathcal{I}(G)$  is the family of all independent sets in a suitably defined undirected version of the graph and  $\Delta_i$  are the suboptimality gaps. The algorithm combines ideas from the EXP3++ algorithm for stochastic and adversarial bandits and the EXP3.G algorithm for feedback graphs with a novel exploration scheme. The scheme, which exploits the structure of the graph to reduce exploration, is key to obtain best-of-both-worlds guarantees with feedback graphs. We also extend our algorithm and results to a setting where the feedback graphs are allowed to change over time.

## [Quantum Algorithms for Sampling Log-Concave Distributions and Estimating Normalizing Constants](#)

- Andrew M. Childs · Tongyang Li · Jin-Peng Liu · Chunhao Wang · Ruizhe Zhang
- abstract@[open-review](#): Given a convex function  $R^d \rightarrow R$ , the problem of sampling from a distribution  $\propto e^{-f(x)}$  is called log-concave sampling. This task has wide applications in machine learning, physics, statistics, etc. In this work, we develop quantum algorithms for sampling log-concave distributions and for estimating their normalizing constants  $\int_{\mathbb{R}^d} e^{-f(x)} dx$ . First, we use underdamped Langevin diffusion to develop quantum algorithms that match the query complexity (in terms of the condition number  $\kappa$  and dimension  $d$ ) of analogous classical algorithms that use gradient (first-order) queries, even though the quantum algorithms use only evaluation (zeroth-order) queries. For estimating normalizing constants, these algorithms also achieve quadratic speedup in the multiplicative error  $\epsilon$ . Second, we develop quantum Metropolis-adjusted Langevin algorithms with query complexity  $\widetilde{O}(\kappa^{1/2} d)$  and  $\widetilde{O}(\kappa^{1/2} d^{3/2} \epsilon)$  for log-concave sampling and normalizing constant estimation, respectively, achieving polynomial speedups in  $\kappa, d, \epsilon$  over the best known classical algorithms by exploiting quantum analogs of the Monte Carlo method and quantum walks. We also prove a  $\Omega(1/\epsilon^{1-o(1)})$  quantum lower bound for estimating normalizing constants, implying near-optimality of our quantum algorithms in  $\epsilon$ .

## [Making Look-Ahead Active Learning Strategies Feasible with Neural Tangent Kernels](#)

- Mohamad Amin Mohamadi · Wonho Bae · Danica J. Sutherland
- abstract@[open-review](#): We propose a new method for approximating active learning acquisition strategies that are based on retraining with hypothetically-labeled candidate data points. Although this is usually infeasible with deep networks, we use the neural tangent kernel to approximate the result of retraining, and prove that this approximation works asymptotically even in an active learning setup -- approximating look-ahead" selection criteria with far less computation required. This also enables us to conduct sequential active learning, i.e.\ updating the model in a streaming regime, without needing to retrain the model with SGD after adding each new data point. Moreover, our querying strategy, which better understands how the model's predictions will change by adding new data points in comparison to the standard ("myopic") criteria, beats other look-ahead strategies by large margins, and achieves equal or better performance compared to state-of-the-art methods on several benchmark datasets in pool-based active learning.

## [Graph Neural Networks are Dynamic Programmers](#)

- Andrew J Dudzik · Petar Veličković
- abstract@[open-review](#): Recent advances in neural algorithmic reasoning with graph neural networks (GNNs) are propped up by the notion of algorithmic alignment. Broadly, a neural network will be better at learning to execute a reasoning task (in terms of sample complexity) if its individual components align well with the target algorithm. Specifically, GNNs are claimed to align with dynamic programming (DP), a general problem-solving strategy which expresses many polynomial-time algorithms. However, has this alignment truly been demonstrated and theoretically quantified? Here we show, using methods from category theory and abstract algebra, that there exists an intricate connection between GNNs and DP, going well beyond the initial observations over individual algorithms such as Bellman-Ford. Exposing this connection, we easily verify several prior findings in the literature, produce better-grounded GNN architectures for edge-centric tasks, and demonstrate empirical results on the CLRS algorithmic reasoning benchmark. We hope our exposition will serve as a foundation for building stronger algorithmically aligned GNNs.

## [Operator Splitting Value Iteration](#)

- Amin Rakhsha · Andrew Wang · Mohammad Ghavamzadeh · Amir-massoud Farahmand
- abstract@[open-review](#): We introduce new planning and reinforcement learning algorithms for discounted MDPs that can utilize an approximate model of the environment to accelerate the convergence of the value function. Inspired by the splitting approach in numerical linear algebra, we introduce Operator Splitting Value Iteration (OS-VI) for both policy evaluation and control problems. OS-VI achieves a much faster convergence rate when the model is accurate enough. We also introduce a sample-based version of the algorithm called OS-Dyna. Unlike the traditional Dyna architecture, OS-Dyna still converges to the correct value function in presence of model approximation error. To the best of our knowledge, this is the only model-based algorithm with this property.

## [Chain of Thought Imitation with Procedure Cloning](#)

- Mengjiao (Sherry) Yang · Dale Schuurmans · Pieter Abbeel · Ofir Nachum
- abstract@[open-review](#): Imitation learning aims to extract high-performance policies from logged demonstrations of expert behavior. It is common to frame imitation learning as a supervised learning problem in which one fits a function approximator to the input-output mapping exhibited by the logged demonstrations (input observations to output actions). While the framing of imitation learning as a supervised input-output learning problem allows for applicability in a wide variety of settings, it is also an overly simplistic view of the problem in situations where the expert demonstrations provide much richer insight into expert behavior. For example, applications such as path navigation, robot manipulation, and strategy games acquire expert demonstrations via planning, search, or some other multi-step algorithm, revealing not just the output action to be imitated but also the procedure for how to determine this action. While these intermediate computations may use tools not available to the agent during inference (e.g., environment simulators), they are nevertheless informative as a way to explain an expert's mapping of state to actions. To properly leverage expert procedure information without relying on the privileged tools the expert may have used to perform the procedure, we propose procedure cloning, which applies supervised sequence prediction to imitate the complete series of expert computations. This way, procedure cloning learns not only what to do (i.e., the output action), but how and why to do it (i.e., the procedure). Through empirical analysis on navigation, simulated robotic manipulation, and game-playing environments, we show that imitating the intermediate computations of an expert's behavior enables procedure cloning to learn policies exhibiting significant generalization to unseen environment configurations, including those configurations for which running the expert's procedure directly is infeasible.

## [Invertible Monotone Operators for Normalizing Flows](#)

- Byeongkeun Ahn · Chiyoon Kim · Youngjoon Hong · Hyunwoo Kim
- abstract@[open-review](#): Normalizing flows model probability distributions by learning invertible transformations that transfer a simple distribution into complex distributions. Since the architecture of ResNet-based normalizing flows is more flexible than that of coupling-based models, ResNet-based normalizing flows have been widely studied in recent years. Despite their architectural flexibility, it is well-known that the current ResNet-based models suffer from constrained Lipschitz constants. In this paper, we propose the monotone formulation to overcome the issue of the Lipschitz constants using monotone operators and provide an in-depth theoretical analysis. Furthermore, we construct an activation function called Concatenated Pila (CPila) to improve gradient flow. The resulting model, Monotone Flows, exhibits an excellent density estimation performance and outperforms existing state-of-the-art normalizing flow models on multiple density estimation benchmarks (MNIST, CIFAR-10, ImageNet32, ImageNet64).

## [Multi-Game Decision Transformers](#)

- Kuang-Huei Lee · Ofir Nachum · Mengjiao (Sherry) Yang · Lisa Lee · Daniel Freeman · Sergio Guadarrama · Ian Fischer · Winnie Xu · Eric Jang · Henryk Michalewski · Igor Mordatch
- abstract@[open-review](#): A longstanding goal of the field of AI is a strategy for compiling diverse experience into a highly capable, generalist agent. In the subfields of vision and language, this was largely achieved by scaling up transformer-based models and training them on large, diverse datasets. Motivated by this progress, we investigate whether the same strategy can be used to produce generalist reinforcement learning agents. Specifically, we show that a single transformer-based model "with a single set of weights" trained purely offline can play a suite of up to 46 Atari games simultaneously at close-to-human performance. When trained and evaluated appropriately, we find that the same trends observed in language and vision hold, including scaling of performance with model size and rapid adaptation to new games via fine-tuning. We compare several approaches in this multi-game setting, such as online and offline RL methods and behavioral cloning, and find that our Multi-Game Decision Transformer models offer the best scalability and performance. We release the pre-trained models and code to encourage further research in this direction.

## [Riemannian Diffusion Models](#)

- Chin-Wei Huang · Milad Aghajohari · Joey Bose · Prakash Panangaden · Aaron Courville
- abstract@[open-review](#): Diffusion models are recent state-of-the-art methods for image generation and likelihood estimation. In this work, we generalize continuous-time diffusion models to arbitrary Riemannian manifolds and derive a variational framework for likelihood estimation. Computationally, we propose new methods for computing the Riemannian divergence which is needed in the likelihood estimation. Moreover, in generalizing the Euclidean case, we prove that maximizing this variational lower-bound is equivalent to Riemannian score matching. Empirically, we demonstrate the expressive power of Riemannian diffusion models on a wide spectrum of smooth manifolds, such as spheres, tori, hyperboloids, and orthogonal groups. Our proposed method achieves new state-of-the-art likelihoods on all benchmarks.

## [Training with More Confidence: Mitigating Injected and Natural Backdoors During Training](#)

- Zhenting Wang · Hailun Ding · Juan Zhai · Shiqing Ma
- abstract@[open-review](#): The backdoor or Trojan attack is a severe threat to deep neural networks (DNNs). Researchers find that DNNs trained on benign data and settings can also learn backdoor behaviors, which is known as the natural backdoor. Existing works on anti-backdoor learning are based on weak observations that the backdoor and benign behaviors can differentiate during training. An adaptive attack with slow poisoning can bypass such defenses. Moreover, these methods cannot defend natural backdoors. We found the fundamental differences between backdoor-related neurons and benign neurons: backdoor-related neurons form a hyperplane as the classification surface across input domains of all affected labels. By further analyzing the training

process and model architectures, we found that piece-wise linear functions cause this hyperplane surface. In this paper, we design a novel training method that forces the training to avoid generating such hyperplanes and thus remove the injected backdoors. Our extensive experiments on five datasets against five state-of-the-art attacks and also benign training show that our method can outperform existing state-of-the-art defenses. On average, the ASR (attack success rate) of the models trained with \sys is 54.83 times lower than undefended models under standard poisoning backdoor attack and 1.75 times lower under the natural backdoor attack. Our code is available at <https://anonymous.4open.science/r/NOLE-84C3>.

## [Vision GNN: An Image is Worth Graph of Nodes](#)

- Kai Han · Yunhe Wang · Jianyuan Guo · Yehui Tang · Enhua Wu
- abstract@[open-review](#): Network architecture plays a key role in the deep learning-based computer vision system. The widely-used convolutional neural network and transformer treat the image as a grid or sequence structure, which is not flexible to capture irregular and complex objects. In this paper, we propose to represent the image as a graph structure and introduce a new \emph{Vision GNN} (ViG) architecture to extract graph-level feature for visual tasks. We first split the image to a number of patches which are viewed as nodes, and construct a graph by connecting the nearest neighbors. Based on the graph representation of images, we build our ViG model to transform and exchange information among all the nodes. ViG consists of two basic modules: Grapher module with graph convolution for aggregating and updating graph information, and FFN module with two linear layers for node feature transformation. Both isotropic and pyramid architectures of ViG are built with different model sizes. Extensive experiments on image recognition and object detection tasks demonstrate the superiority of our ViG architecture. We hope this pioneering study of GNN on general visual tasks will provide useful inspiration and experience for future research.

## [GhostNetV2: Enhance Cheap Operation with Long-Range Attention](#)

- Yehui Tang · Kai Han · Jianyuan Guo · Chang Xu · Chao Xu · Yunhe Wang
- abstract@[open-review](#): Light-weight convolutional neural networks (CNNs) are specially designed for applications on mobile devices with faster inference speed yet modest performance. The convolutional operation can only capture local information in a window region, which prevents performance from being further improved. Introducing self-attention into convolution can capture global information well, but it will largely encumber the actual speed. In this paper, we propose a hardware-friendly attention mechanism (dubbed DFC attention) and then present a new GhostNetV2 architecture for mobile applications. The proposed DFC attention is constructed based on fully-connected layers, which can not only execute fast on common hardware but also capture the dependence between long-range pixels. We further revisit the expressiveness bottleneck in previous GhostNet and propose to enhance expanded features produced by cheap operations with DFC attention, so that a GhostNetV2 block can aggregate local and long-range information simultaneously. Extensive experiments demonstrate the superiority of GhostNetV2 over existing architectures. For example, it achieves 75.3% top-1 accuracy on ImageNet with 167M FLOPs, significantly suppressing GhostNetV1 (74.5%) with a similar computational cost.

## [Chefs' Random Tables: Non-Trigonometric Random Features](#)

- Valerii Likhoshevstov · Krzysztof M Choromanski · Kumar Avinava Dubey · Frederick Liu · Tamas Sarlos · Adrian Weller
- abstract@[open-review](#): We introduce chefs' random tables (CRTs), a new class of non-trigonometric random features (RFs) to approximate Gaussian and softmax kernels. CRTs are an alternative to standard random kitchen sink (RKS) methods, which inherently rely on the trigonometric maps. We present variants of CRTs where RFs are positive, a key requirement for applications in recent low-rank Transformers. Further variance reduction is possible by leveraging statistics which are simple to compute. One instantiation of CRTs, the optimal positive random features (OPRFs), is to our knowledge the first RF method for unbiased softmax kernel estimation with positive and bounded RFs, resulting in exponentially small tails and much lower variance than its counterparts. As we show, orthogonal random features applied in OPRFs provide additional variance reduction for any dimensionality \$d\$ (not only asymptotically for sufficiently large \$d\$, as for RKS). We test CRTs on many tasks ranging from non-parametric classification to training Transformers for text, speech and image data, obtaining new state-of-the-art results for low-rank text Transformers, while providing linear space and time complexity.

## [Redistribution of Weights and Activations for AdderNet Quantization](#)

- Ying Nie · Kai Han · Haikang Diao · Chuanjian Liu · Enhua Wu · Yunhe Wang
- abstract@[open-review](#): Adder Neural Network (AdderNet) provides a new way for developing energy-efficient neural networks by replacing the expensive multiplications in convolution with cheaper additions (i.e., L1-norm). To achieve higher hardware efficiency, it is necessary to further study the low-bit quantization of AdderNet. Due to the limitation that the commutative law in multiplication does not hold in L1-norm, the well-established quantization methods on convolutional networks cannot be applied on AdderNets. Thus, the existing AdderNet quantization techniques propose to use only one shared scale to quantize both the weights and activations simultaneously. Admittedly, such an approach can keep the commutative law in the L1-norm quantization process, while the accuracy drop after low-bit quantization cannot be ignored. To this end, we first thoroughly analyze the difference on distributions of weights and activations in AdderNet and then propose a new quantization algorithm by redistributing the weights and the activations. Specifically, the pre-trained full-precision weights in different kernels are clustered into different groups, then the intra-group sharing and inter-group independent scales can be adopted. To further compensate the accuracy drop caused by the distribution difference, we then develop a lossless range clamp scheme for weights and a simple yet effective outliers clamp strategy for activations. Thus, the functionality of full-precision weights and the representation ability of full-precision activations can be fully preserved. The effectiveness of the proposed quantization method for AdderNet is well-verified on several benchmarks, e.g., our 4-bit post-training quantized adder ResNet-18 achieves an 66.5% top-1 accuracy on the ImageNet with comparable energy efficiency, which is about 8.5% higher than that of the previous AdderNet quantization methods.

## [Hierarchical Agglomerative Graph Clustering in Poly-Logarithmic Depth](#)

- Laxman Dhulipala · David Eisenstat · Jakub Lacki · Vahab Mirrokni · Jessica Shi
- abstract@[open-review](#): Obtaining scalable algorithms for \emph{hierarchical agglomerative clustering} (HAC) is of significant interest due to the massive size of real-world datasets. At the same time, efficiently parallelizing HAC is difficult due to the seemingly sequential nature of the algorithm. In this paper, we address this issue and present ParHAC, the first efficient parallel HAC algorithm with sublinear depth for the widely-used average-linkage function. In particular, we provide a \$(1+\epsilon)\$-approximation algorithm for this problem on \$m\$ edge graphs using \$\tilde{O}(m)\$ work and poly-logarithmic depth. Moreover, we show that obtaining similar bounds for \emph{exact} average-linkage HAC is not possible under standard complexity-theoretic assumptions. We complement our theoretical results with a comprehensive study of the ParHAC algorithm in terms of its scalability, performance, and quality, and compare with several state-of-the-art sequential and parallel baselines. On a broad set of large publicly-available real-world datasets, we find that ParHAC obtains a 50.1x speedup on average over the best sequential baseline, while achieving quality similar to the exact HAC algorithm. We also show that ParHAC can cluster one of the largest publicly available graph datasets with 124 billion edges in a little over three hours using a commodity multicore machine.

## [A Fully Transformer-Based Object Detector with Fine-Coarse Crossing Representations](#)

- Zhishan Li · Ying Nie · Kai Han · Jianyuan Guo · Lei Xie · Yunhe Wang
- abstract@[open-review](#): Transformer-based object detectors have shown competitive performance recently. Compared with convolutional neural networks limited by the relatively small receptive fields, the advantage of transformer for visual tasks is the capacity to perceive long-range dependencies among all image patches, while the deficiency is that the local fine-grained information is not fully excavated. In this paper, we introduce the Fine-grained and Coarse-grained crossing representations for building an efficient Detection Transformer (FCDT). Specifically, we propose a local-global cross fusion

module to establish the connection between local fine-grained features and global coarse-grained features. Besides, we propose a Fine-Coarse Aware Neck which enables det tokens to interact with both coarse-grained and fine-grained features. Furthermore, we present an efficient feature integration module to fuse multi-scale representations from different stages. Experimental results on Microsoft COCO dataset demonstrate the effectiveness of our approach. For instance, our FCDT model achieves 48.1 AP with 173G FLOPs, which possesses higher accuracy and less computation compared with the state-of-the-art fully transformer-based detector ViDT.

## [Efficient Non-Parametric Optimizer Search for Diverse Tasks](#)

- Ruochen Wang · Yuanhao Xiong · Minhao Cheng · Cho-Jui Hsieh
- abstract@[open-review](#): Efficient and automated design of optimizers plays a crucial role in full-stack AutoML systems. However, prior methods in optimizer search are often limited by their scalability, generability, or sample efficiency. With the goal of democratizing research and application of optimizer search, we present the first efficient, scalable and generalizable framework that can directly search on the tasks of interest. We first observe that optimizer updates are fundamentally mathematical expressions applied to the gradient. Inspired by the innate tree structure of the underlying math expressions, we re-arrange the space of optimizers into a super-tree, where each path encodes an optimizer. This way, optimizer search can be naturally formulated as a path-finding problem, allowing a variety of well-established tree traversal methods to be used as the search algorithm. We adopt an adaptation of the Monte Carlo method to tree search, equipped with rejection sampling and equivalent-form detection that leverage the characteristics of optimizer update rules to further boost the sample efficiency. We provide a diverse set of tasks to benchmark our algorithm and demonstrate that, with only 128 evaluations, the proposed framework can discover optimizers that surpass both human-designed counterparts and prior optimizer search methods.

## [Factored DRO: Factored Distributionally Robust Policies for Contextual Bandits](#)

- Tong Mu · Yash Chandak · Tatsunori Hashimoto · Emma Brunskill
- abstract@[open-review](#): While there has been extensive work on learning from offline data for contextual multi-armed bandit settings, existing methods typically assume there is no environment shift: that the learned policy will operate in the same environmental process as that of data collection. However, this assumption may overly limit the use of these methods for many practical situations where there may be distribution shifts. In this work we propose Factored Distributionally Robust Optimization (Factored-DRO), which is able to separately handle distribution shifts in the context distribution and shifts in the reward generating process. Prior work that either ignores potential shifts in the context, or considers them jointly, can lead to performance that is too conservative and does not consider context shift at all, especially under certain forms of reward feedback. Our Factored-DRO objective mitigates this by considering the shifts separately, and our proposed estimators are consistent and converge asymptotically. We also introduce a practical algorithm and demonstrate promising empirical results in environments based on real-world datasets, such as voting outcomes and scene classification.

## [Positive-Unlabeled Learning using Random Forests via Recursive Greedy Risk Minimization](#)

- Jonathan Wilton · Nan Ye · Miao Xu · Abigail Koay · Ryan Ko
- abstract@[open-review](#): The need to learn from positive and unlabeled data, or PU learning, arises in many applications and has attracted increasing interest. While random forests are known to perform well on many tasks with positive and negative data, recent PU algorithms are generally based on deep neural networks, and the potential of tree-based PU learning is under-explored. In this paper, we propose new random forest algorithms for PU-learning. Key to our approach is a new interpretation of decision tree algorithms for positive and negative data as \{recursive greedy risk minimization algorithms\}. We extend this perspective to the PU setting to develop new decision tree learning algorithms that directly minimizes PU-data based estimators for the expected risk. This allows us to develop an efficient PU random forest algorithm, PU extra trees. Our approach features three desirable properties: it is robust to the choice of the loss function in the sense that various loss functions lead to the same decision trees; it requires little hyperparameter tuning as compared to neural network based PU learning; it supports a feature importance that directly measures a feature's contribution to risk minimization. Our algorithms demonstrate strong performance on several datasets. Our code is available at \url{https://github.com/puetpaper/PUEExtraTrees}.

## [Scalable Interpretability via Polynomials](#)

- Abhimanyu Dubey · Filip Radenovic · Dhruv Mahajan
- abstract@[open-review](#): Generalized Additive Models (GAMs) have quickly become the leading choice for fully-interpretable machine learning. However, unlike uninterpretable methods such as DNNs, they lack expressive power and easy scalability, and are hence not a feasible alternative for real-world tasks. We present a new class of GAMs that use tensor rank decompositions of polynomials to learn powerful, \{em fully-interpretable\} models. Our approach, titled Scalable Polynomial Additive Models (SPAM) is effortlessly scalable and models \{em all\} higher-order feature interactions without a combinatorial parameter explosion. SPAM outperforms all current interpretable approaches, and matches DNN/XGBoost performance on a series of real-world benchmarks with up to hundreds of thousands of features. We demonstrate by human subject evaluations that SPAMs are demonstrably more interpretable in practice, and are hence an effortless replacement for DNNs for creating interpretable and high-performance systems suitable for large-scale machine learning.

## [Diffusion Models as Plug-and-Play Priors](#)

- Alexandros Graikos · Nikolay Malkin · Nebojsa Jojic · Dimitris Samaras
- abstract@[open-review](#): We consider the problem of inferring high-dimensional data \$x\$ in a model that consists of a prior \$p(x)\$ and an auxiliary constraint \$c(x,y)\$. In this paper, the prior is an independently trained denoising diffusion generative model. The auxiliary constraint is expected to have a differentiable form, but can come from diverse sources. The possibility of such inference turns diffusion models into plug-and-play modules, thereby allowing a range of potential applications in adapting models to new domains and tasks, such as conditional generation or image segmentation. The structure of diffusion models allows us to perform approximate inference by iterating differentiation through the fixed denoising network enriched with different amounts of noise at each step. Considering many noised versions of \$x\$ in evaluation of its fitness is a novel search mechanism that may lead to new algorithms for solving combinatorial optimization problems.

## [Non-Gaussian Tensor Programs](#)

- Eugene Golikov · Greg Yang
- abstract@[open-review](#): The Tensor Programs framework has produced a series of powerful results by 1) expressing any deep learning computation of concern as a principled composition of element-wise nonlinearities and matrix multiplication, and 2) inductively reasoning about the program behavior as the sizes of the matrices in the program tend to infinity. For example, this framework helped to show that infinitely wide neural networks exhibit Gaussian process behavior at initialization and evolve like a kernel model during training in the so-called NTK parameterization (Yang, 2019b, 2020a; Yang and Littwin, 2021). Moreover, this framework yielded a novel parameterization, coined  $\hat{\mathcal{P}}$  (Yang and Hu, 2021), that for the first time enabled hyperparameter tuning for enormous networks too expensive to train more than once (Yang et al., 2022). However, this framework has so far been limited to Gaussian initialized weights, while uniform or truncated Gaussian distributions are more prevalent in practice. This work extends Tensor Programs to general non-Gaussian weights, thus recovering all of the above results in all practical settings.

## [When to Update Your Model: Constrained Model-based Reinforcement Learning](#)

- Tianying Ji · Yu Luo · Fuchun Sun · Mingxuan Jing · Fengxiang He · Wenbing Huang
- abstract@[open-review](#): Designing and analyzing model-based RL (MBRL) algorithms with guaranteed monotonic improvement has been challenging, mainly due to the interdependence between policy optimization and model learning. Existing discrepancy bounds generally ignore the impacts of model shifts, and their corresponding algorithms are prone to degrade performance by drastic model updating. In this work, we first propose a novel and general theoretical scheme for a non-decreasing performance guarantee of MBRL. Our follow-up derived bounds reveal the relationship between model shifts and performance improvement. These discoveries encourage us to formulate a constrained lower-bound optimization problem to permit the monotonicity of MBRL. A further example demonstrates that learning models from a dynamically-varying number of explorations benefit the eventual returns. Motivated by these analyses, we design a simple but effective algorithm CMLO (Constrained Model-shift Lower-bound Optimization), by introducing an event-triggered mechanism that flexibly determines when to update the model. Experiments show that CMLO surpasses other state-of-the-art methods and produces a boost when various policy optimization methods are employed.

## [Sampling without Replacement Leads to Faster Rates in Finite-Sum Minimax Optimization](#)

- Aniket Das · Bernhard Schölkopf · Michael Muehlebach
- abstract@[open-review](#): We analyze the convergence rates of stochastic gradient algorithms for smooth finite-sum minimax optimization and show that, for many such algorithms, sampling the data points \{without replacement\} leads to faster convergence compared to sampling with replacement. For the smooth and strongly convex-strongly concave setting, we consider gradient descent ascent and the proximal point method, and present a unified analysis of two popular without-replacement sampling strategies, namely \{Random Reshuffling\} (RR), which shuffles the data every epoch, and \{Single Shuffling\} or \{Shuffle Once\} (SO), which shuffles only at the beginning. We obtain tight convergence rates for RR and SO and demonstrate that these strategies lead to faster convergence than uniform sampling. Moving beyond convexity, we obtain similar results for smooth nonconvex-nonconcave objectives satisfying a two-sided Polyak-\L{}ojasiewicz inequality. Finally, we demonstrate that our techniques are general enough to analyze the effect of \{data-ordering attacks\}, where an adversary manipulates the order in which data points are supplied to the optimizer. Our analysis also recovers tight rates for the \{incremental gradient\} method, where the data points are not shuffled at all.

## [Trustworthy Monte Carlo](#)

- Juha Harvainen · Mikko Koivisto · Petteri Kaski
- abstract@[open-review](#): Monte Carlo integration is a key technique for designing randomized approximation schemes for counting problems, with applications, e.g., in machine learning and statistical physics. The technique typically enables massively parallel computation, however, with the risk that some of the delegated computations contain spontaneous or adversarial errors. We present an orchestration of the computations such that the outcome is accompanied with a proof of correctness. Specifically, we adopt an algebraic proof system developed in computational complexity theory, in which the proof is represented by a polynomial; evaluating the polynomial at a random point amounts to a verification of the proof with probabilistic guarantees. We give examples of known Monte Carlo estimators that admit verifiable extensions with moderate computational overhead: for the permanent of zero-one matrices, for the model count of disjunctive normal form formulas, and for the gradient of logistic regression models. We also discuss the prospects and challenges of engineering efficient verifiable approximation schemes more generally.

## [On Deep Generative Models for Approximation and Estimation of Distributions on Manifolds](#)

- Biraj Dahal · Alexander Havrilla · Minshuo Chen · Tuo Zhao · Wenjing Liao
- abstract@[open-review](#): Deep generative models have experienced great empirical successes in distribution learning. Many existing experiments have demonstrated that deep generative networks can efficiently generate high-dimensional complex data from a low-dimensional easy-to-sample distribution. However, this phenomenon can not be justified by existing theories. The widely held manifold hypothesis speculates that real-world data sets, such as natural images and signals, exhibit low-dimensional geometric structures. In this paper, we take such low-dimensional data structures into consideration by assuming that data distributions are supported on a low-dimensional manifold. We prove approximation and estimation theories of deep generative networks for estimating distributions on a low-dimensional manifold under the Wasserstein-1 loss. We show that the Wasserstein-1 loss converges to zero at a fast rate depending on the intrinsic dimension instead of the ambient data dimension. Our theory leverages the low-dimensional geometric structures in data sets and justifies the practical power of deep generative models. We require no smoothness assumptions on the data distribution which is desirable in practice.

## [Instance-Dependent Policy Learning for Linear MDPs via Online Experiment Design](#)

- Andrew Wagenmaker · Kevin Jamieson
- abstract@[open-review](#): While much progress has been made in understanding the minimax sample complexity of reinforcement learning (RL)---the complexity of learning on the worst-case" instance---such measures of complexity often do not capture the true difficulty of learning. In practice, on aneasy" instance, we might hope to achieve a complexity far better than that achievable on the worst-case instance. In this work we seek to understand this instance-dependent" complexity of learning in the setting of RL with linear function approximation. We propose an algorithm, PEDEL, which achieves a fine-grained instance-dependent measure of complexity, the first of its kind in the RL with function approximation setting, thereby capturing the difficulty of learning on each particular problem instance. Through an explicit example, we show that PEDEL yields provable gains over low-regret, minimax-optimal algorithms and that such algorithms are unable to hit the instance-optimal rate. Our approach relies on a novel online experiment design-based procedure which focuses the exploration budget on the directions" most relevant to learning a near-optimal policy, and may be of independent interest.

## [Concept Activation Regions: A Generalized Framework For Concept-Based Explanations](#)

- Jonathan Crabb · Mihaela van der Schaar
- abstract@[open-review](#): Concept-based explanations permit to understand the predictions of a deep neural network (DNN) through the lens of concepts specified by users. Existing methods assume that the examples illustrating a concept are mapped in a fixed direction of the DNN's latent space. When this holds true, the concept can be represented by a concept activation vector (CAV) pointing in that direction. In this work, we propose to relax this assumption by allowing concept examples to be scattered across different clusters in the DNN's latent space. Each concept is then represented by a region of the DNN's latent space that includes these clusters and that we call concept activation region (CAR). To formalize this idea, we introduce an extension of the CAV formalism that is based on the kernel trick and support vector classifiers. This CAR formalism yields global concept-based explanations and local concept-based feature importance. We prove that CAR explanations built with radial kernels are invariant under latent space isometries. In this way, CAR assigns the same explanations to latent spaces that have the same geometry. We further demonstrate empirically that CARs offer (1) more accurate descriptions of how concepts are scattered in the DNN's latent space; (2) global explanations that are closer to human concept annotations and (3) concept-based feature importance that meaningfully relate concepts with each other. Finally, we use CARs to show that DNNs can autonomously rediscover known scientific concepts, such as the prostate cancer grading system.

## [Learning to Navigate Wikipedia with Graph Diffusion Models](#)

- Manzil Zaheer · Kenneth Marino · Will Grathwohl · John Schultz · Wenling Shang · Sheila Babayan · Arun Ahuja · Ishita Dasgupta · Christine Kaeser-Chen · Rob Fergus
- abstract@[open-review](#): A fundamental ability of an intelligent web-based agent is seeking out and acquiring new information. Internet search engines reliably find the correct vicinity but the top results may be a few links away from the desired target. A complementary approach is navigation via hyperlinks, employing a policy that comprehends local content and selects a link that moves it closer to the target. In this paper we show that using behavioral cloning of randomly sampled trajectories is sufficient to learn an effective link selection policy. We demonstrate the approach on a graph version of Wikipedia with \$38\\$M nodes and \$387\\$M edges. The model is able to efficiently navigate between nodes \$5\\$ and \$20\\$ steps apart \$96\%\$ and \$92\%\$ of the time, respectively. We then use the resulting embeddings and policy in a downstream fact verification task where, in combination with basic TF-IDF search and ranking methods, they are able to obtain competitive results to state-of-the-art methods.

## [ZIN: When and How to Learn Invariance by Environment Inference?](#)

- Yong Lin · Shengyu Zhu · Lu Tan · Peng Cui
- abstract@[open-review](#): It is commonplace to encounter heterogeneous data, of which some aspects of the data distribution may vary but the underlying causal mechanisms remain constant. When data are divided into distinct environments according to the heterogeneity, recent invariant learning methods have proposed to learn robust and invariant models based on this environment partition. It is hence tempting to utilize the inherent heterogeneity even when environment partition is not provided. Unfortunately, in this work, we show that learning invariant features under this circumstance is fundamentally impossible without further inductive biases or additional information. Then, we propose a framework to jointly learn environment partition and invariant representation, assisted by additional auxiliary information. We derive sufficient and necessary conditions for our framework to provably identify invariant features under a fairly general setting. Experimental results on both synthetic and real world datasets validate our analysis and demonstrate an improved performance of the proposed framework over existing methods. Finally, our results also raise the need of making the role of inductive biases more explicit in future works, when considering learning invariant models without environment partition.

## [Information-Theoretic Generative Model Compression with Variational Energy-based Model](#)

- Minsoo Kang · Hyewon Yoo · Eunhee Kang · Sehwan Ki · Hyong Euk Lee · Bohyung Han
- abstract@[open-review](#): We propose an information-theoretic knowledge distillation approach for the compression of generative adversarial networks, which aims to maximize the mutual information between the teacher and student networks via a variational optimization based on an energy-based model. Because the direct computation of the mutual information in continuous domains is intractable, our approach alternatively optimizes the student network by maximizing the variational lower bound on the mutual information. To achieve a tight lower bound, we introduce an energy-based model relying on a deep neural network to represent a flexible variational distribution that deals with high-dimensional images and consider spatial dependencies between pixels, effectively. Since the proposed method is a generic optimization algorithm, it can be conveniently incorporated into arbitrary generative adversarial networks and even dense prediction networks, e.g., image enhancement models. We demonstrate that the proposed algorithm achieves outstanding performance in model compression of generative adversarial networks consistently when combined with several existing models.

## [Understanding Programmatic Weak Supervision via Source-aware Influence Function](#)

- Jieyu Zhang · Haonan Wang · Cheng-Yu Hsieh · Alexander Ratner
- abstract@[open-review](#): Programmatic Weak Supervision (PWS) aggregates the source votes of multiple weak supervision sources into probabilistic training labels, which are in turn used to train an end model. With its increasing popularity, it is critical to have some tool for users to understand the influence of each component (e.g., the source vote or training data) in the pipeline and interpret the end model behavior. To achieve this, we build on Influence Function (IF) and propose source-aware IF, which leverages the generation process of the probabilistic labels to decompose the end model's training objective and then calculate the influence associated with each (data, source, class) tuple. These primitive influence score can then be used to estimate the influence of individual component of PWS, such as source vote, supervision source, and training data. On datasets of diverse domains, we demonstrate multiple use cases: (1) interpreting incorrect predictions from multiple angles that reveals insights for debugging the PWS pipeline, (2) identifying mislabeling of sources with a gain of 9%-37% over baselines, and (3) improving the end model's generalization performance by removing harmful components in the training objective (13%-24% better than ordinary IF).

## [Detecting Abrupt Changes in Sequential Pairwise Comparison Data](#)

- Wanshan Li · Alessandro Rinaldo · Daren Wang
- abstract@[open-review](#): The Bradley-Terry-Luce (BTL) model is a classic and very popular statistical approach for eliciting a global ranking among a collection of items using pairwise comparison data. In applications in which the comparison outcomes are observed as a time series, it is often the case that data are non-stationary, in the sense that the true underlying ranking changes over time. In this paper we are concerned with localizing the change points in a high-dimensional BTL model with piece-wise constant parameters. We propose novel and practicable algorithms based on dynamic programming that can consistently estimate the unknown locations of the change points. We provide consistency rates for our methodology that depend explicitly on the model parameters, the temporal spacing between two consecutive change points and the magnitude of the change. We corroborate our findings with extensive numerical experiments and a real-life example.

## [Archimedes Meets Privacy: On Privately Estimating Quantiles in High Dimensions Under Minimal Assumptions](#)

- Omri Ben-Eliezer · Dan Mikulincer · Ilias Zadik
- abstract@[open-review](#): The last few years have seen a surge of work on high dimensional statistics under privacy constraints, mostly following two main lines of work: the "worst case" line, which does not make any distributional assumptions on the input data; and the "strong assumptions" line, which assumes that the data is generated from specific families, e.g., subgaussian distributions. In this work we take a middle ground, obtaining new differentially private algorithms with polynomial sample complexity for estimating quantiles in high-dimensions, as well as estimating and sampling points of high Tukey depth, all working under very mild distributional assumptions. From the technical perspective, our work relies upon deep robustness results in the convex geometry literature, demonstrating how such results can be used in a private context. Our main object of interest is the (convex) floating body (FB), a notion going back to Archimedes, which is a robust and well studied high-dimensional analogue of the interquantile range of a distribution. We show how one can privately, and with polynomially many samples, (a) output an approximate interior point of the FB -- e.g., "a typical user" in a high-dimensional database -- by leveraging the robustness of the Steiner point of the FB; and at the expense of polynomially many more samples, (b) produce an approximate uniform sample from the FB, by constructing a private noisy projection oracle.

## [UMIX: Improving Importance Weighting for Subpopulation Shift via Uncertainty-Aware Mixup](#)

- Zongbo Han · Zhipeng Liang · Fan Yang · Liu Liu · Lanqing Li · Yatao Bian · Peilin Zhao · Bingzhe Wu · Changqing Zhang · Jianhua Yao
- abstract@[open-review](#): Subpopulation shift wildly exists in many real-world machine learning applications, referring to the training and test distributions containing the same subpopulation groups but varying in subpopulation frequencies. Importance reweighting is a normal way to handle the subpopulation shift issue by imposing constant or adaptive sampling weights on each sample in the training dataset. However, some recent studies have recognized that most of these approaches fail to improve the performance over empirical risk minimization especially when applied to over-parameterized neural networks. In this work, we propose a simple yet practical framework, called uncertainty-aware mixup (UMIX), to mitigate the overfitting issue in over-parameterized models by reweighting the "mixed" samples according to the sample uncertainty. The training-trajectories-based uncertainty estimation is

equipped in the proposed UMIX for each sample to flexibly characterize the subpopulation distribution. We also provide insightful theoretical analysis to verify that UMIX achieves better generalization bounds over prior works. Further, we conduct extensive empirical studies across a wide range of tasks to validate the effectiveness of our method both qualitatively and quantitatively.

## [Where2comm: Communication-Efficient Collaborative Perception via Spatial Confidence Maps](#)

- Yue Hu · Shaoheng Fang · Zixing Lei · Yiqi Zhong · Siheng Chen
- abstract@[open-review](#): Multi-agent collaborative perception could significantly upgrade the perception performance by enabling agents to share complementary information with each other through communication. It inevitably results in a fundamental trade-off between perception performance and communication bandwidth. To tackle this bottleneck issue, we propose a spatial confidence map, which reflects the spatial heterogeneity of perceptual information. It empowers agents to only share spatially sparse, yet perceptually critical information, contributing to where to communicate. Based on this novel spatial confidence map, we propose Where2comm, a communication-efficient collaborative perception framework. Where2comm has two distinct advantages: i) it considers pragmatic compression and uses less communication to achieve higher perception performance by focusing on perceptually critical areas; and ii) it can handle varying communication bandwidth by dynamically adjusting spatial areas involved in communication. To evaluate Where2comm, we consider 3D object detection in both real-world and simulation scenarios with two modalities (camera/LiDAR) and two agent types (cars/drones) on four datasets: OPV2V, V2X-Sim, DAIR-V2X, and our original CoPerception-UAVs. Where2comm consistently outperforms previous methods; for example, it achieves more than \$100,000 \times lower communication volume and still outperforms DiscoNet and V2X-ViT on OPV2V. Our code is available at \url{https://github.com/MediaBrain-SJTU/where2comm}.

## [LISA: Learning Interpretable Skill Abstractions from Language](#)

- Divyansh Garg · Skanda Vaidyanath · Kuno Kim · Jiaming Song · Stefano Ermon
- abstract@[open-review](#): Learning policies that effectually utilize language instructions in complex, multi-task environments is an important problem in imitation learning. While it is possible to condition on the entire language instruction directly, such an approach could suffer from generalization issues. To encode complex instructions into skills that can generalize to unseen instructions, we propose Learning Interpretable Skill Abstractions (LISA), a hierarchical imitation learning framework that can learn diverse, interpretable skills from language-conditioned demonstrations. LISA uses vector quantization to learn discrete skill codes that are highly correlated with language instructions and the behavior of the learned policy. In navigation and robotic manipulation environments, LISA is able to outperform a strong non-hierarchical baseline in the low data regime and compose learned skills to solve tasks containing unseen long-range instructions. Our method demonstrates a more natural way to condition on language in sequential decision-making problems and achieve interpretable and controllable behavior with the learned skills.

## [Natural Color Fool: Towards Boosting Black-box Unrestricted Attacks](#)

- Shengming Yuan · Qilong Zhang · Lianli Gao · Yaya Cheng · Jingkuan Song
- abstract@[open-review](#): Unrestricted color attacks, which manipulate semantically meaningful color of an image, have shown their stealthiness and success in fooling both human eyes and deep neural networks. However, current works usually sacrifice the flexibility of the uncontrolled setting to ensure the naturalness of adversarial examples. As a result, the black-box attack performance of these methods is limited. To boost transferability of adversarial examples without damaging image quality, we propose a novel Natural Color Fool (NCF) which is guided by realistic color distributions sampled from a publicly available dataset and optimized by our neighborhood search and initialization reset. By conducting extensive experiments and visualizations, we convincingly demonstrate the effectiveness of our proposed method. Notably, on average, results show that our NCF can outperform state-of-the-art approaches by \textbf{15.0\%}\sim\textbf{32.9\%} for fooling normally trained models and \textbf{10.0\%}\sim\textbf{25.3\%} for evading defense methods. Our code is available at \url{https://github.com/ylhz/Natural-Color-Fool}.

## [AutoST: Towards the Universal Modeling of Spatio-temporal Sequences](#)

- Jianxin Li · Shuai Zhang · Hui Xiong · Haoyi Zhou
- abstract@[open-review](#): The analysis of spatio-temporal sequences plays an important role in many real-world applications, demanding a high model capacity to capture the interdependence among spatial and temporal dimensions. Previous studies provided separated network design in three categories: spatial first, temporal first, and spatio-temporal synchronous. However, the manually-designed heterogeneous models can hardly meet the spatio-temporal dependency capturing priority for various tasks. To address this, we proposed a universal modeling framework with three distinctive characteristics: (i) Attention-based network backbone, including S2T Layer (spatial first), T2S Layer (temporal first), and STS Layer (spatio-temporal synchronous). (ii) The universal modeling framework, named UniST, with a unified architecture that enables flexible modeling priorities with the proposed three different modules. (iii) An automatic search strategy, named AutoST, automatically searches the optimal spatio-temporal modeling priority by network architecture search. Extensive experiments on five real-world datasets demonstrate that UniST with any single type of our three proposed modules can achieve state-of-the-art performance. Furthermore, AutoST can achieve overwhelming performance with UniST.

## [On Solving Class Incremental Learning in Continual Learning](#)

- Gyuhak Kim · Changnan Xiao · Tatsuya Konishi · Zixuan Ke · Bing Liu
- abstract@[open-review](#): Continual learning (CL) is concerned with learning a sequence of tasks incrementally. There are two popular CL settings, class incremental learning (CIL) and task incremental learning (TIL). A major challenge of CL is catastrophic forgetting (CF). While a number of techniques are already available to effectively overcome CF for TIL, CIL remains to be highly challenging. So far, little study has been done to provide a theoretical guidance on how to solve the CIL problem. This paper performs such a study. It first shows that probabilistically, the CIL problem can be decomposed into two sub-problems: within-task prediction and task-id prediction. It further proves that task-id prediction is correlated to out-of-distribution (OOD) detection, which connects CIL and OOD detection and at the same time, offers a principled approach to solving CIL. Experiments have been conducted to empirically verify the theoretical result. Based on the result, new CIL methods are also designed, which outperform strong baselines by a large margin.

## [Transcormer: Transformer for Sentence Scoring with Sliding Language Modeling](#)

- Kaitao Song · Yichong Leng · Xu Tan · Yicheng Zou · Tao Qin · Dongsheng Li
- abstract@[open-review](#): Sentence scoring aims at measuring the likelihood score of a sentence and is widely used in many natural language processing scenarios, like reranking, which is to select the best sentence from multiple candidates. Previous works on sentence scoring mainly adopted either causal language modeling (CLM) like GPT or masked language modeling (MLM) like BERT, which have some limitations: 1) CLM only utilizes unidirectional information for the probability estimation of a sentence without considering bidirectional context, which affects the scoring quality; 2) MLM can only estimate the probability of partial tokens at a time and thus requires multiple forward passes to estimate the probability of the whole sentence, which incurs large computation and time cost. In this paper, we propose \textit{Transcormer} -- a Transformer model with a novel \textit{sliding language modeling} (SLM) for sentence scoring. Specifically, our SLM adopts a triple-stream self-attention mechanism to estimate the probability of all tokens in a sentence with bidirectional context and only requires a single forward pass. SLM can avoid the limitations of CLM (only unidirectional context) and MLM (multiple forward passes) and inherit their advantages, and thus achieve high effectiveness and efficiency in scoring. Experimental results on multiple tasks demonstrate that our method achieves better performance than other language modelings.

## On the Effectiveness of Persistent Homology

- Renata Turkes · Guido Montufar · Nina Otter
- abstract@[open-review](#): Persistent homology (PH) is one of the most popular methods in Topological Data Analysis. Even though PH has been used in many different types of applications, the reasons behind its success remain elusive; in particular, it is not known for which classes of problems it is most effective, or to what extent it can detect geometric or topological features. The goal of this work is to identify some types of problems where PH performs well or even better than other state-of-the-art methods in data analysis. We consider three fundamental shape analysis tasks: the detection of the number of holes, curvature and convexity from 2D and 3D point clouds sampled from shapes. Experiments demonstrate that PH is successful in these tasks, outperforming several baselines, including PointNet, an architecture inspired precisely by the properties of point clouds. In addition, we observe that PH remains effective for limited computational resources and limited training data, as well as out-of-distribution test data, including various data transformations and noise. For convexity detection, we provide a theoretical guarantee that PH is effective for this task, and demonstrate the detection of a convexity measure on the FLAVIA dataset of plant leaf images.

## Trading off Utility, Informativeness, and Complexity in Emergent Communication

- Mycal Tucker · Roger Levy · Julie Shah · Noga Zaslavsky
- abstract@[open-review](#): Emergent communication research often focuses on optimizing task-specific utility as a driver for communication. However, there is substantial evidence that human languages evolve under pressure to efficiently compress meanings into communication signals by optimizing the Information Bottleneck tradeoff between informativeness and complexity. In this work, we study how trading off these three factors --- utility, informativeness, and complexity --- shapes emergent communication, including compared to human communication. To this end, we propose Vector-Quantized Variational Information Bottleneck (VQ-VIB), a method for training neural agents to compress inputs into discrete signals embedded in a continuous space. We train agents via VQ-VIB and compare their performance to previously proposed neural architectures, both in multi-agent reinforcement learning settings and in a Lewis reference game. Across all neural architectures and settings, taking into account communicative informativeness benefits communication convergence rates, and penalizing communicative complexity leads to human-like lexicon sizes while maintaining high utility. Additionally, we find that VQ-VIB outperforms other discrete communication methods. This work demonstrates how fundamental principles that are believed to characterize human language evolution may inform emergent communication in artificial agents.

## Grounded Video Situation Recognition

- Zeeshan Khan · C.V. Jawahar · Makarand Tapaswi
- abstract@[open-review](#): Dense video understanding requires answering several questions such as \emph{who is doing what to whom, with what, how, why, and where}. Recently, Video Situation recognition (VidSitu) is framed as a task for structured prediction of multiple events, their relationships, their actions and various verb-role pairs attached to descriptive entities. This task poses several challenges in identifying, disambiguating, and co-referencing entities across multiple verb-role pairs, but also faces some challenges of evaluation. In this work, we propose to add spatio-temporal grounding as an essential component of the structured prediction task and present a novel three stage Transformer model, VideoWhisperer, that is empowered to make joint predictions. In stage one, we learn contextualised embeddings for video features in parallel with key objects that appear in the video clips to enable fine-grained spatio-temporal reasoning. The second stage sees verb-role queries attend and pool information from object embeddings, localising answers to questions posed about the action. The final stage generates these answers as captions to describe each verb-role pair present in the video. Our model operates on a group of events (clips) simultaneously and predicts verbs, verb-role pairs, their nouns, and their grounding on-the-fly. When evaluated on a grounding-augmented version of the VidSitu dataset, we observe a large improvement in entity captioning accuracy, as well as the ability to localize verb-roles without grounding annotations at training time.

## HierSpeech: Bridging the Gap between Text and Speech by Hierarchical Variational Inference using Self-supervised Representations for Speech Synthesis

- Sang-Hoon Lee · Seung-Bin Kim · Ji-Hyun Lee · Eunwoo Song · Min-Jae Hwang · Seong-Whan Lee
- abstract@[open-review](#): This paper presents HierSpeech, a high-quality end-to-end text-to-speech (TTS) system based on a hierarchical conditional variational autoencoder (VAE) utilizing self-supervised speech representations. Recently, single-stage TTS systems, which directly generate raw speech waveform from text, have been getting interest thanks to their ability in generating high-quality audio within a fully end-to-end training pipeline. However, there is still a room for improvement in the conventional TTS systems. Since it is challenging to infer both the linguistic and acoustic attributes from the text directly, missing the details of attributes, specifically linguistic information, is inevitable, which results in mispronunciation and over-smoothing problem in their synthetic speech. To address the aforementioned problem, we leverage self-supervised speech representations as additional linguistic representations to bridge an information gap between text and speech. Then, the hierarchical conditional VAE is adopted to connect these representations and to learn each attribute hierarchically by improving the linguistic capability in latent representations. Compared with the state-of-the-art TTS system, HierSpeech achieves +0.303 comparative mean opinion score, and reduces the phoneme error rate of synthesized speech from 9.16% to 5.78% on the VCTK dataset. Furthermore, we extend our model to HierSpeech-U, an untranscribed text-to-speech system. Specifically, HierSpeech-U can adapt to a novel speaker by utilizing self-supervised speech representations without text transcripts. The experimental results reveal that our method outperforms publicly available TTS models, and show the effectiveness of speaker adaptation with untranscribed speech.

## Structural Kernel Search via Bayesian Optimization and Symbolical Optimal Transport

- Matthias Bitzer · Mona Meister · Christoph Zimmer
- abstract@[open-review](#): Despite recent advances in automated machine learning, model selection is still a complex and computationally intensive process. For Gaussian processes (GPs), selecting the kernel is a crucial task, often done manually by the expert. Additionally, evaluating the model selection criteria for Gaussian processes typically scales cubically in the sample size, rendering kernel search particularly computationally expensive. We propose a novel, efficient search method through a general, structured kernel space. Previous methods solved this task via Bayesian optimization and relied on measuring the distance between GP's directly in function space to construct a kernel-kernel. We present an alternative approach by defining a kernel-kernel over the symbolic representation of the statistical hypothesis that is associated with a kernel. We empirically show that this leads to a computationally more efficient way of searching through a discrete kernel space.

## Theory and Approximate Solvers for Branched Optimal Transport with Multiple Sources

- Peter Lippmann · Enrique Fita Sanmartín · Fred Hamprecht
- abstract@[open-review](#): Branched Optimal Transport (BOT) is a generalization of optimal transport in which transportation costs along an edge are subadditive. This subadditivity models an increase in transport efficiency when shipping mass along the same route, favoring branched transportation networks. We here study the NP-hard optimization of BOT networks connecting a finite number of sources and sinks in  $\mathbb{R}^2$ . First, we show how to efficiently find the best geometry of a BOT network for many sources and sinks, given a topology. Second, we argue that a topology with more than three edges meeting at a branching point is never optimal. Third, we show that the results obtained for the Euclidean plane generalize directly to optimal transportation networks on two-dimensional Riemannian manifolds. Finally, we present a simple but effective approximate BOT solver combining geometric optimization with a combinatorial optimization of the network topology.

## Understanding Robust Learning through the Lens of Representation Similarities

- Christian Cianfarani · Arjun Nitin Bhagoji · Vikash Sehwag · Ben Zhao · Prateek Mittal · Heather Zheng
- abstract@[open-review](#): Representation learning, \textit{i.e.} the generation of representations useful for downstream applications, is a task of fundamental importance that underlies much of the success of deep neural networks (DNNs). Recently, \emph{robustness to adversarial examples} has emerged as a desirable property for DNNs, spurring the development of robust training methods that account for adversarial examples. In this paper, we aim to understand how the properties of representations learned by robust training differ from those obtained from standard, non-robust training. This is critical to diagnosing numerous salient pitfalls in robust networks, such as, degradation of performance on benign inputs, poor generalization of robustness, and increase in over-fitting. We utilize a powerful set of tools known as representation similarity metrics, across 3 vision datasets, to obtain layer-wise comparisons between robust and non-robust DNNs with different architectures, training procedures and adversarial constraints. Our experiments highlight hitherto unseen properties of robust representations that we posit underlie the behavioral differences of robust networks. We discover a lack of specialization in robust networks' representations along with a disappearance of `block structure'. We also find overfitting during robust training largely impacts deeper layers. These, along with other findings, suggest ways forward for the design and training of better robust networks.

## Revisiting Heterophily For Graph Neural Networks

- Sitao Luan · Chenqing Hua · Qincheng Lu · Jiaqi Zhu · Mingde Zhao · Shuyuan Zhang · Xiao-Wen Chang · Doina Precup
- abstract@[open-review](#): Graph Neural Networks (GNNs) extend basic Neural Networks (NNs) by using graph structures based on the relational inductive bias (homophily assumption). While GNNs have been commonly believed to outperform NNs in real-world tasks, recent work has identified a non-trivial set of datasets where their performance compared to NNs is not satisfactory. Heterophily has been considered the main cause of this empirical observation and numerous works have been put forward to address it. In this paper, we first revisit the widely used homophily metrics and point out that their consideration of only graph-label consistency is a shortcoming. Then, we study heterophily from the perspective of post-aggregation node similarity and define new homophily metrics, which are potentially advantageous compared to existing ones. Based on this investigation, we prove that some harmful cases of heterophily can be effectively addressed by local diversification operation. Then, we propose the Adaptive Channel Mixing (ACM), a framework to adaptively exploit aggregation, diversification and identity channels to extract richer localized information in each baseline GNN layer. ACM is more powerful than the commonly used uni-channel framework for node classification tasks on heterophilic graphs. When evaluated on 10 benchmark node classification tasks, ACM-augmented baselines consistently achieve significant performance gain, exceeding state-of-the-art GNNs on most tasks without incurring significant computational burden.

## A Characterization of Semi-Supervised Adversarially Robust PAC Learnability

- Idan Attias · Steve Hanneke · Yishay Mansour
- abstract@[open-review](#): We study the problem of learning an adversarially robust predictor to test time attacks in the semi-supervised PAC model. We address the question of how many labeled and unlabeled examples are required to ensure learning. We show that having enough unlabeled data (the size of a labeled sample that a fully-supervised method would require), the labeled sample complexity can be arbitrarily smaller compared to previous works, and is sharply characterized by a different complexity measure. We prove nearly matching upper and lower bounds on this sample complexity. This shows that there is a significant benefit in semi-supervised robust learning even in the worst-case distribution-free model, and establishes a gap between supervised and semi-supervised label complexities which is known not to hold in standard non-robust PAC learning.

## Delving into Sequential Patches for Deepfake Detection

- Jiazhi Guan · Hang Zhou · Zhibin Hong · Errui Ding · Jingdong Wang · Chengbin Quan · Youjian Zhao
- abstract@[open-review](#): Recent advances in face forgery techniques produce nearly visually untraceable deepfake videos, which could be harmful to society with malicious intentions. To tackle this problem, researches have been devoted to deepfake detection. While the importance of local low-level cues and temporal information has been verified in this task, previous methods struggle in achieving generalizability across deepfake methods and robustness towards image post-processings. In this work, we propose the Local- & Temporal-aware Transformer-based Deepfake Detection (\$\textbf{LTTD}\$) framework, which adopts a local-to-global learning protocol with a particular focus on the valuable temporal information within local sequences. Specifically, we propose a Local Sequence Transformer (LST), which models the temporal consistency on sequences of restricted spatial regions, where low-level information is hierarchically enhanced with shallow layers of learned 3D filters. Based on the local temporal embeddings, we then achieve the final classification in a global contrastive way. Extensive experiments on popular datasets validate that our approach effectively spots local forgery cues and achieves state-of-the-art performance.

## Probabilistic Transformer: Modelling Ambiguities and Distributions for RNA Folding and Molecule Design

- JÃ¶rg Franke · Frederic Runge · Frank Hutter
- abstract@[open-review](#): Our world is ambiguous and this is reflected in the data we use to train our algorithms. This is especially true when we try to model natural processes where collected data is affected by noisy measurements and differences in measurement techniques. Sometimes, the process itself can be ambiguous, such as in the case of RNA folding, where a single nucleotide sequence can fold into multiple structures that occur with different probabilities. This ambiguity suggests that a predictive model should have similar probabilistic characteristics to match the data it models. Therefore, we propose a hierarchical latent distribution to enhance one of the most successful deep learning models, the Transformer, to accommodate these sorts of ambiguities and data distributions. We show the benefits of our approach by learning the hidden distribution on a synthetic task, with state-of-the-art results in RNA folding when training on highly ambiguous data, capable of reconstructing structure distributions and demonstrate its generative capabilities on property-based molecule design by implicitly learning the underlying property distributions and outperforming existing work.

## Untargeted Backdoor Watermark: Towards Harmless and Stealthy Dataset Copyright Protection

- Yiming Li · Yang Bai · Yong Jiang · Yong Yang · Shu-Tao Xia · Bo Li
- abstract@[open-review](#): Deep neural networks (DNNs) have demonstrated their superiority in practice. Arguably, the rapid development of DNNs is largely benefited from high-quality (open-sourced) datasets, based on which researchers and developers can easily evaluate and improve their learning methods. Since the data collection is usually time-consuming or even expensive, how to protect their copyrights is of great significance and worth further exploration. In this paper, we revisit dataset ownership verification. We find that existing verification methods introduced new security risks in DNNs trained on the protected dataset, due to the targeted nature of poison-only backdoor watermarks. To alleviate this problem, in this work, we explore the untargeted backdoor watermarking scheme, where the abnormal model behaviors are not deterministic. Specifically, we introduce two dispersibilities and prove their correlation, based on which we design the untargeted backdoor watermark under both poisoned-label and clean-label settings. We also discuss how to use the proposed untargeted backdoor watermark for dataset ownership verification. Experiments on benchmark datasets verify the effectiveness of our methods and their resistance to existing backdoor defenses.

## Synergy-of-Experts: Collaborate to Improve Adversarial Robustness

- Sen Cui · Jingfeng ZHANG · Jian Liang · Bo Han · Masashi Sugiyama · Changshui Zhang

- abstract@[open-review](#): Learning adversarially robust models require invariant predictions to a small neighborhood of its natural inputs, often encountering insufficient model capacity. There is research showing that learning multiple sub-models in an ensemble could mitigate this insufficiency, further improving the generalization and the robustness. However, the ensemble's voting-based strategy excludes the possibility that the true predictions remain with the minority. Therefore, this paper further improves the ensemble through a collaboration scheme---Synergy-of-Experts (SoE). Compared with the voting-based strategy, the SoE enables the possibility of correct predictions even if there exists a single correct sub-model. In SoE, every sub-model fits its specific vulnerability area and reserves the rest of the sub-models to fit other vulnerability areas, which effectively optimizes the utilization of the model capacity. Empirical experiments verify that SoE outperforms various ensemble methods against white-box and transfer-based adversarial attacks.

## [HSDF: Hybrid Sign and Distance Field for Modeling Surfaces with Arbitrary Topologies](#)

- Li Wang · jie Yang · Weikai Chen · Xiaoxu Meng · Bo Yang · Jintao Li · Lin Gao
- abstract@[open-review](#): Neural implicit function based on signed distance field (SDF) has achieved impressive progress in reconstructing 3D models with high fidelity. However, such approaches can only represent closed shapes. Recent works based on unsigned distance function (UDF) are proposed to handle both watertight and open surfaces. Nonetheless, as UDF is signless, its direct output is limited to point cloud, which imposes an additional challenge on extracting high-quality meshes from discrete points. To address this issue, we present a new learnable implicit representation, coded HSDF, that connects the good ends of SDF and UDF. In particular, HSDF is able to represent arbitrary topologies containing both closed and open surfaces while being compatible with existing iso-surface extraction techniques for easy field-to-mesh conversion. In addition to predicting a UDF, we propose to learn an additional sign field via a simple classifier. Unlike traditional SDF, HSDF is able to locate the surface of interest before level surface extraction by generating surface points following NDF~\cite{chibane2020ndf}. We are then able to obtain open surfaces via an adaptive meshing approach that only instantiates regions containing surface into a polygon mesh. We also propose HSDF-Net, a dedicated learning framework that factorizes the learning of HSDF into two easier problems. Experiments on multiple datasets show that HSDF outperforms state-of-the-art techniques both qualitatively and quantitatively.

## [Private Isotonic Regression](#)

- Badih Ghazi · Pritish Kamath · Ravi Kumar · Pasin Manurangsi
- abstract@[open-review](#): In this paper we consider the problem of differentially private (DP) algorithms for isotonic regression. For the most general problem of isotonic regression over a partially ordered set (poset)  $\mathcal{X}$  and for any Lipschitz loss function, we obtain a pure-DP algorithm that, given  $n$  input points, has an expected excess empirical risk of roughly  $\mathrm{width}(\mathcal{X}) \cdot \log|\mathcal{X}| / n$ , where  $\mathrm{width}(\mathcal{X})$  is the width of the poset. In contrast, we also obtain a near-matching lower bound of roughly  $(\mathrm{width}(\mathcal{X}) + \log |\mathcal{X}|) / n$ , that holds even for approximate-DP algorithms. Moreover, we show that the above bounds are essentially the best that can be obtained without utilizing any further structure of the poset. In the special case of a totally ordered set and for  $\ell_1$  and  $\ell_2$  losses, our algorithm can be implemented in near-linear running time; we also provide extensions of this algorithm to the problem of private isotonic regression with additional structural constraints on the output function.

## [Semi-Supervised Semantic Segmentation via Gentle Teaching Assistant](#)

- Ying Jin · Jiaqi Wang · Dahua Lin
- abstract@[open-review](#): Semi-Supervised Semantic Segmentation aims at training the segmentation model with limited labeled data and a large amount of unlabeled data. To effectively leverage the unlabeled data, pseudo labeling, along with the teacher-student framework, is widely adopted in semi-supervised semantic segmentation. Though proved to be effective, this paradigm suffers from incorrect pseudo labels which inevitably exist and are taken as auxiliary training data. To alleviate the negative impact of incorrect pseudo labels, we delve into the current Semi-Supervised Semantic Segmentation frameworks. We argue that the unlabeled data with pseudo labels can facilitate the learning of representative features in the feature extractor, but it is unreliable to supervise the mask predictor. Motivated by this consideration, we propose a novel framework, Gentle Teaching Assistant (GTA-Seg) to disentangle the effects of pseudo labels on feature extractor and mask predictor of the student model. Specifically, in addition to the original teacher-student framework, our method introduces a teaching assistant network which directly learns from pseudo labels generated by the teacher network. The gentle teaching assistant (GTA) is coined gentle since it only transfers the beneficial feature representation knowledge in the feature extractor to the student model in an Exponential Moving Average (EMA) manner, protecting the student model from the negative influences caused by unreliable pseudo labels in the mask predictor. The student model is also supervised by reliable labeled data to train an accurate mask predictor, further facilitating feature representation. Extensive experiment results on benchmark datasets validate that our method shows competitive performance against previous methods. We promise to release our code towards reproducibility.

## [A Scalable Deterministic Global Optimization Algorithm for Training Optimal Decision Tree](#)

- Kaixun Hua · Jiayang Ren · Yankai Cao
- abstract@[open-review](#): The training of optimal decision tree via mixed-integer programming (MIP) has attracted much attention in recent literature. However, for large datasets, state-of-the-art approaches struggle to solve the optimal decision tree training problems to a provable global optimal solution within a reasonable time. In this paper, we reformulate the optimal decision tree training problem as a two-stage optimization problem and propose a tailed reduced-space branch and bound algorithm to train optimal decision tree for the classification tasks with continuous features. We present several structure-exploiting lower and upper bounding methods. The computation of bounds can be decomposed into the solution of many small-scale subproblems and can be naturally parallelized. With these bounding methods, we prove that our algorithm can converge by branching only on variables representing the optimal decision tree structure, which is invariant to the size of datasets. Moreover, we propose a novel sample reduction method that can predetermine the cost of part of samples at each BB node. Combining the sample reduction method with the parallelized bounding strategies, our algorithm can be extremely scalable and find global optimal solutions with a small optimality gap for datasets with over 245,000 samples. We test 21 real-world datasets from UCI Repository. The results reveal that for datasets with over 7,000 samples, our algorithm can, on average, improve the training accuracy by 3.2% and testing accuracy by 2.8%, compared to the current state-of-the-art.

## [Decoupled Self-supervised Learning for Non-Homophilous Graphs](#)

- Teng Xiao · Zhengyu Chen · Zhimeng Guo · Zeyang Zhuang · Suhang Wang
- abstract@[open-review](#): In this paper, we study the problem of conducting self-supervised learning for node representation learning on non-homophilous graphs. Existing self-supervised learning methods typically assume the graph is homophilous where linked nodes often belong to the same class or have similar features. However, such assumptions of homophily do not always hold true in real-world graphs. We address this problem by developing a decoupled self-supervised learning (DSSL) framework for graph neural networks. DSSL imitates a generative process of nodes and links from latent variable modeling of the semantic structure, which decouples different underlying semantics between different neighborhoods into the self-supervised node learning process. Our DSSL framework is agnostic to the encoders and does not need prefabricated augmentations, thus is flexible to different graphs. To effectively optimize the framework with latent variables, we derive the evidence lower-bound of the self-supervised objective and develop a scalable training algorithm with variational inference. We also provide a theoretical analysis that justifies it enjoys the better downstream performance. Extensive experiments on various types of non-homophilous benchmarks demonstrate that our proposed framework can significantly achieve better performance compared with competitive self-supervised learning baselines.

## Optimistic Mirror Descent Either Converges to Nash or to Strong Coarse Correlated Equilibria in Bimatrix Games

- Ioannis Anagnostides · Gabriele Farina · Ioannis Panageas · Tuomas Sandholm
- abstract@[open-review](#): We show that, for any sufficiently small fixed  $\epsilon > 0$ , when both players in a general-sum two-player (bimatrix) game employ optimistic mirror descent (OMD) with smooth regularization, learning rate  $\eta = O(\epsilon^2)$  and  $T = \Omega(\text{poly}(1/\epsilon))$  repetitions, either the dynamics reach an  $\epsilon$ -approximate Nash equilibrium (NE), or the average correlated distribution of play is an  $\epsilon$ -strong coarse correlated equilibrium (CCE): any possible unilateral deviation does not only leave the player worse, but will decrease its utility by  $\Omega(\text{poly}(\epsilon))$ . As an immediate consequence, when the iterates of OMD are bounded away from being Nash equilibria in a bimatrix game, we guarantee convergence to an exact CCE after only  $O(1)$  iterations. Our results reveal that uncoupled no-regret learning algorithms can converge to CCE in general-sum games remarkably faster than to NE in, for example, zero-sum games. To establish this, we show that when OMD does not reach arbitrarily close to a NE, the cumulative regret of both players is not only negative, but decays linearly with time. Given that regret is the canonical measure of performance in online learning, our results suggest that cycling behavior of no-regret learning algorithms in games can be justified in terms of efficiency.

## Wasserstein K-means for clustering probability distributions

- Yubo Zhuang · Xiaohui Chen · Yun Yang
- abstract@[open-review](#): Clustering is an important exploratory data analysis technique to group objects based on their similarity. The widely used K-means clustering method relies on some notion of distance to partition data into a fewer number of groups. In the Euclidean space, centroid-based and distance-based formulations of the K-means are equivalent. In modern machine learning applications, data often arise as probability distributions and a natural generalization to handle measure-valued data is to use the optimal transport metric. Due to non-negative Alexandrov curvature of the Wasserstein space, barycenters suffer from regularity and non-robustness issues. The peculiar behaviors of Wasserstein barycenters may make the centroid-based formulation fail to represent the within-cluster data points, while the more direct distance-based K-means approach and its semidefinite program (SDP) relaxation are capable of recovering the true cluster labels. In the special case of clustering Gaussian distributions, we show that the SDP relaxed Wasserstein K-means can achieve exact recovery given the clusters are well-separated under the  $2$ -Wasserstein metric. Our simulation and real data examples also demonstrate that distance-based K-means can achieve better classification performance over the standard centroid-based K-means for clustering probability distributions and images.

## DataMUX: Data Multiplexing for Neural Networks

- Vishvak Murahari · Carlos Jimenez · Runzhe Yang · Karthik Narasimhan
- abstract@[open-review](#): In this paper, we introduce data multiplexing (DataMUX), a technique that enables deep neural networks to process multiple inputs simultaneously using a single compact representation. DataMUX demonstrates that neural networks are capable of generating accurate predictions over mixtures of inputs, resulting in increased inference throughput with minimal extra memory requirements. Our approach uses two key components -- 1) a multiplexing layer that performs a fixed linear transformation to each input before combining them to create a "mixed" representation of the same size as a single input, which is then processed by the base network, and 2) a demultiplexing layer that converts the base network's output back into independent representations before producing predictions for each input. We show the viability of DataMUX for different architectures (Transformers, and to a much lesser extent MLPs and CNNs) across six different tasks spanning sentence classification, named entity recognition and image classification. For instance, DataMUX for Transformers can multiplex up to 20x/40x inputs, achieving up to 11x/18x increase in inference throughput with absolute performance drops of  $<2\%$  and  $<4\%$  respectively compared to a vanilla Transformer on MNLI, a natural language inference task. We also provide a theoretical construction for multiplexing in self-attention networks and analyze the effect of various design elements in DataMUX.

## Pre-trained Adversarial Perturbations

- Yuanhao Ban · Yinpeng Dong
- abstract@[open-review](#): Self-supervised pre-training has drawn increasing attention in recent years due to its superior performance on numerous downstream tasks after fine-tuning. However, it is well-known that deep learning models lack the robustness to adversarial examples, which can also invoke security issues to pre-trained models, despite being less explored. In this paper, we delve into the robustness of pre-trained models by introducing Pre-trained Adversarial Perturbations (PAPs), which are universal perturbations crafted for the pre-trained models to maintain the effectiveness when attacking fine-tuned ones without any knowledge of the downstream tasks. To this end, we propose a Low-Level Layer Lifting Attack (L4A) method to generate effective PAPs by lifting the neuron activations of low-level layers of the pre-trained models. Equipped with an enhanced noise augmentation strategy, L4A is effective at generating more transferable PAPs against the fine-tuned models. Extensive experiments on typical pre-trained vision models and ten downstream tasks demonstrate that our method improves the attack success rate by a large margin compared to the state-of-the-art methods.

## Finding Optimal Arms in Non-stochastic Combinatorial Bandits with Semi-bandit Feedback and Finite Budget

- Jasmin Brandt · Björn Haddenhorst · Viktor Bengs · Eyke Höllemeier
- abstract@[open-review](#): We consider the combinatorial bandits problem with semi-bandit feedback under finite sampling budget constraints, in which the learner can carry out its action only for a limited number of times specified by an overall budget. The action is to choose a set of arms, whereupon feedback for each arm in the chosen set is received. Unlike existing works, we study this problem in a non-stochastic setting with subset-dependent feedback, i.e., the semi-bandit feedback received could be generated by an oblivious adversary and also might depend on the chosen set of arms. In addition, we consider a general feedback scenario covering both the numerical-based as well as preference-based case and introduce a sound theoretical framework for this setting guaranteeing sensible notions of optimal arms, which a learner seeks to find. We suggest a generic algorithm suitable to cover the full spectrum of conceivable arm elimination strategies from aggressive to conservative. Theoretical questions about the sufficient and necessary budget of the algorithm to find the best arm are answered and complemented by deriving lower bounds for any learning algorithm for this problem scenario.

## Saliency-Aware Neural Architecture Search

- Ramtin Hosseini · Pengtao Xie
- abstract@[open-review](#): Recently a wide variety of NAS methods have been proposed and achieved considerable success in automatically identifying highly-performing architectures of neural networks for the sake of reducing the reliance on human experts. Existing NAS methods ignore the fact that different input data elements (e.g., image pixels) have different importance (or saliency) in determining the prediction outcome. They treat all data elements as being equally important and therefore lead to suboptimal performance. To address this problem, we propose an end-to-end framework which dynamically detects saliency of input data, reweights data using saliency maps, and searches architectures on saliency-reweighted data. Our framework is based on four-level optimization, which performs four learning stages in a unified way. At the first stage, a model is trained with its architecture tentatively fixed. At the second stage, saliency maps are generated using the trained model. At the third stage, the model is retrained on saliency-reweighted data. At the fourth stage, the model is evaluated on a validation set and the architecture is updated by minimizing the validation loss. Experiments on several datasets demonstrate the effectiveness of our framework.

## ESCADA: Efficient Safety and Context Aware Dose Allocation for Precision Medicine

- Ilker Demirel · Ahmet Alparslan Celik · Cem Tekin
- abstract@[open-review](#): Finding an optimal individualized treatment regimen is considered one of the most challenging precision medicine problems. Various patient characteristics influence the response to the treatment, and hence, there is no one-size-fits-all regimen. Moreover, the administration of an unsafe dose during the treatment can have adverse effects on health. Therefore, a treatment model must ensure patient safety while efficiently optimizing the course of therapy. We study a prevalent medical problem where the treatment aims to keep a physiological variable in a safe range and preferably close to a target level, which we refer to as leveling. Such a task may be relevant in numerous other domains as well. We propose ESCADA, a novel and generic multi-armed bandit (MAB) algorithm tailored for the leveling task, to make safe, personalized, and context-aware dose recommendations. We derive high probability upper bounds on its cumulative regret and safety guarantees. Following ESCADA's design, we also describe its Thompson sampling-based counterpart. We discuss why the straightforward adaptations of the classical MAB algorithms such as GP-UCB may not be a good fit for the leveling task. Finally, we make *in silico* experiments on the bolus-insulin dose allocation problem in type-1 diabetes mellitus disease and compare our algorithms against the famous GP-UCB algorithm, the rule-based dose calculators, and a clinician.

## [On the Robustness of Deep Clustering Models: Adversarial Attacks and Defenses](#)

- Anshuman Chhabra · Ashwin Sekhari · Prasant Mohapatra
- abstract@[open-review](#): Clustering models constitute a class of unsupervised machine learning methods which are used in a number of application pipelines, and play a vital role in modern data science. With recent advancements in deep learning-- deep clustering models have emerged as the current state-of-the-art over traditional clustering approaches, especially for high-dimensional image datasets. While traditional clustering approaches have been analyzed from a robustness perspective, no prior work has investigated adversarial attacks and robustness for deep clustering models in a principled manner. To bridge this gap, we propose a blackbox attack using Generative Adversarial Networks (GANs) where the adversary does not know which deep clustering model is being used, but can query it for outputs. We analyze our attack against multiple state-of-the-art deep clustering models and real-world datasets, and find that it is highly successful. We then employ some natural unsupervised defense approaches, but find that these are unable to mitigate our attack. Finally, we attack Face++, a production-level face clustering API service, and find that we can significantly reduce its performance as well. Through this work, we thus aim to motivate the need for truly robust deep clustering models.

## [Learning Active Camera for Multi-Object Navigation](#)

- Peihao Chen · Dongyu Ji · Kunyang Lin · Weiwen Hu · Wenbing Huang · Thomas Li · Mingkui Tan · Chuang Gan
- abstract@[open-review](#): Getting robots to navigate to multiple objects autonomously is essential yet difficult in robot applications. One of the key challenges is how to explore environments efficiently with camera sensors only. Existing navigation methods mainly focus on fixed cameras and few attempts have been made to navigate with active cameras. As a result, the agent may take a very long time to perceive the environment due to limited camera scope. In contrast, humans typically gain a larger field of view by looking around for a better perception of the environment. How to make robots perceive the environment as efficiently as humans is a fundamental problem in robotics. In this paper, we consider navigating to multiple objects more efficiently with active cameras. Specifically, we cast moving camera to a Markov Decision Process and reformulate the active camera problem as a reinforcement learning problem. However, we have to address two new challenges: 1) how to learn a good camera policy in complex environments and 2) how to coordinate it with the navigation policy. To address these, we carefully design a reward function to encourage the agent to explore more areas by moving camera actively. Moreover, we exploit human experience to infer a rule-based camera action to guide the learning process. Last, to better coordinate two kinds of policies, the camera policy takes navigation actions into account when making camera moving decisions. Experimental results show our camera policy consistently improves the performance of multi-object navigation over four baselines on two datasets.

## [On Sample Optimality in Personalized Collaborative and Federated Learning](#)

- Mathieu Even · Laurent Massoulié · Kevin Scaman
- abstract@[open-review](#): In personalized federated learning, each member of a potentially large set of agents aims to train a model minimizing its loss function averaged over its local data distribution. We study this problem under the lens of stochastic optimization, focusing on a scenario with a large number of agents, that each possess very few data samples from their local data distribution. Specifically, we prove novel matching lower and upper bounds on the number of samples required from all agents to approximately minimize the generalization error of a fixed agent. We provide strategies matching these lower bounds, based on a gradient filtering approach: given prior knowledge on some notion of distance between local data distributions, agents filter and aggregate stochastic gradients received from other agents, in order to achieve an optimal bias-variance trade-off. Finally, we quantify the impact of using rough estimations of the distances between local distributions of agents, based on a very small number of local samples.

## [SPoVT: Semantic-Prototype Variational Transformer for Dense Point Cloud Semantic Completion](#)

- Sheng Yu Huang · Hao-Yu Hsu · Frank Wang
- abstract@[open-review](#): Point cloud completion is an active research topic for 3D vision and has been widely studied in recent years. Instead of directly predicting missing point cloud from the partial input, we introduce a Semantic-Prototype Variational Transformer (SPoVT) in this work, which takes both partial point cloud and their semantic labels as the inputs for semantic point cloud object completion. By observing and attending at geometry and semantic information as input features, our SPoVT would derive point cloud features and their semantic prototypes for completion purposes. As a result, our SPoVT not only performs point cloud completion with varying resolution, it also allows manipulation of different semantic parts of an object. Experiments on benchmark datasets would quantitatively and qualitatively verify the effectiveness and practicality of our proposed model.

## [Debiased Self-Training for Semi-Supervised Learning](#)

- Baixu Chen · Junguang Jiang · Ximei Wang · Pengfei Wan · Jianmin Wang · Mingsheng Long
- abstract@[open-review](#): Deep neural networks achieve remarkable performances on a wide range of tasks with the aid of large-scale labeled datasets. Yet these datasets are time-consuming and labor-exhaustive to obtain on realistic tasks. To mitigate the requirement for labeled data, self-training is widely used in semi-supervised learning by iteratively assigning pseudo labels to unlabeled samples. Despite its popularity, self-training is well-known to be unreliable and often leads to training instability. Our experimental studies further reveal that the bias in semi-supervised learning arises from both the problem itself and the inappropriate training with potentially incorrect pseudo labels, which accumulates the error in the iterative self-training process. To reduce the above bias, we propose Debiased Self-Training (DST). First, the generation and utilization of pseudo labels are decoupled by two parameter-independent classifier heads to avoid direct error accumulation. Second, we estimate the worst case of self-training bias, where the pseudo labeling function is accurate on labeled samples, yet makes as many mistakes as possible on unlabeled samples. We then adversarially optimize the representations to improve the quality of pseudo labels by avoiding the worst case. Extensive experiments justify that DST achieves an average improvement of 6.3% against state-of-the-art methods on standard semi-supervised learning benchmark datasets and 18.9% against FixMatch on 13 diverse tasks. Furthermore, DST can be seamlessly adapted to other self-training methods and help stabilize their training and balance performance across classes in both cases of training from scratch and finetuning from pre-trained models.

## [Disentangling the Predictive Variance of Deep Ensembles through the Neural Tangent Kernel](#)

- Seijin Kobayashi · Pau Vilimelis Aceituno · Johannes von Oswald
- abstract@[open-review](#): Identifying unfamiliar inputs, also known as out-of-distribution (OOD) detection, is a crucial property of any decision making process. A simple and empirically validated technique is based on deep ensembles where the variance of predictions over different neural networks acts as

a substitute for input uncertainty. Nevertheless, a theoretical understanding of the inductive biases leading to the performance of deep ensemble's uncertainty estimation is missing. To improve our description of their behavior, we study deep ensembles with large layer widths operating in simplified linear training regimes, in which the functions trained with gradient descent can be described by the neural tangent kernel. We identify two sources of noise, each inducing a distinct inductive bias in the predictive variance. We further show theoretically and empirically that both noise sources affect the predictive variance of non-linear deep ensembles in toy models and realistic settings. Finally, we propose practical ways to eliminate possibly unfavorable noise sources leading to improved OOD detection in deep ensembles.

## [Hierarchical Channel-spatial Encoding for Communication-efficient Collaborative Learning](#)

- Qihua ZHOU · Song Guo · YI LIU · Jie Zhang · Jiewei Zhang · Tao GUO · Zhenda XU · Zhihao Qu
- abstract@[open-review](#): It witnesses that the collaborative learning (CL) systems often face the performance bottleneck of limited bandwidth, where multiple low-end devices continuously generate data and transmit intermediate features to the cloud for incremental training. To this end, improving the communication efficiency by reducing traffic size is one of the most crucial issues for realistic deployment. Existing systems mostly compress features at pixel level and ignore the characteristics of feature structure, which could be further exploited for more efficient compression. In this paper, we take new insights into implementing scalable CL systems through a hierarchical compression on features, termed Stripe-wise Group Quantization (SGQ). Different from previous unstructured quantization methods, SGQ captures both channel and spatial similarity in pixels, and simultaneously encodes features in these two levels to gain a much higher compression ratio. In particular, we refactor feature structure based on inter-channel similarity and bound the gradient deviation caused by quantization, in forward and backward passes, respectively. Such a double-stage pipeline makes SGQ hold a sublinear convergence order as the vanilla SGD-based optimization. Extensive experiments show that SGQ achieves a higher traffic reduction ratio by up to 15.97 times and provides 9.22 times image processing speedup over the uniform quantized training, while preserving adequate model accuracy as FP32 does, even using 4-bit quantization. This verifies that SGQ can be applied to a wide spectrum of edge intelligence applications.

## [RenyiCL: Contrastive Representation Learning with Skew Renyi Divergence](#)

- Kyungmin Lee · Jinwoo Shin
- abstract@[open-review](#): Contrastive representation learning seeks to acquire useful representations by estimating the shared information between multiple views of data. Here, the choice of data augmentation is sensitive to the quality of learned representations: as harder the data augmentations are applied, the views share more task-relevant information, but also task-irrelevant one that can hinder the generalization capability of representation. Motivated by this, we present a new robust contrastive learning scheme, coined R\enycl, which can effectively manage harder augmentations by utilizing R\eniy divergence. Our method is built upon the variational lower bound of a Renyi divergence, but a naive usage of a variational method exhibits unstable training due to the large variance. To tackle this challenge, we propose a novel contrastive objective that conducts variational estimation of a skew Renyi divergence and provides a theoretical guarantee on how variational estimation of skew divergence leads to stable training. We show that R\eniy contrastive learning objectives perform innate hard negative sampling and easy positive sampling simultaneously so that it can selectively learn useful features and ignore nuisance features. Through experiments on ImageNet, we show that R\eniy contrastive learning with stronger augmentations outperforms other self-supervised methods without extra regularization or computational overhead. Also, we validate our method on various domains such as graph and tabular datasets, showing empirical gain over original contrastive methods.

## [Probabilistic Missing Value Imputation for Mixed Categorical and Ordered Data](#)

- Yuxuan Zhao · Alex Townsend · Madeleine Udell
- abstract@[open-review](#): Many real-world datasets contain missing entries and mixed data types including categorical and ordered (e.g. continuous and ordinal) variables. Imputing the missing entries is necessary, since many data analysis pipelines require complete data, but challenging especially for mixed data. This paper proposes a probabilistic imputation method using an extended Gaussian copula model that supports both single and multiple imputation. The method models mixed categorical and ordered data using a latent Gaussian distribution. The unordered characteristics of categorical variables is explicitly modeled using the argmax operator. The method makes no assumptions on the data marginals nor does it require tuning any hyperparameters. Experimental results on synthetic and real datasets show that imputation with the extended Gaussian copula outperforms the current state-of-the-art for both categorical and ordered variables in mixed data.

## [TotalSelfScan: Learning Full-body Avatars from Self-Portrait Videos of Faces, Hands, and Bodies](#)

- Junting Dong · Qi Fang · Yudong Guo · Sida Peng · Qing Shuai · Hujun Bao · Xiaowei Zhou
- abstract@[open-review](#): Recent advances in neural implicit functions make it possible to reconstruct a human model from a monocular self-rotation human video. While they present impressive results of the human body, the quality of reconstructed face and hands are relatively low. The main reason is the image region occupied by these parts is very small compared to the body. To solve this problem, we propose TotalSelfScan, which reconstructs the full-body human from several monocular self-rotation videos that focus on the face, hands, and body, respectively. Compared to recording a single body video, this setting has almost no additional cost but provides abundant details of essential parts. To learn the full-body model, instead of encoding the whole body in a single-part network, we propose a novel multi-part representation to model separate parts and then fuse the part-specific observations into the unified human model. Once learned, the human model enables rendering photorealistic free-viewpoint videos under novel human poses. Experiments show that TotalSelfScan can significantly improve the performance on the face and hands compared to the existing methods.

## [Distributional Reward Estimation for Effective Multi-agent Deep Reinforcement Learning](#)

- Jifeng Hu · Yanchao Sun · Hechang Chen · Sili Huang · haiyin piao · Yi Chang · Lichao Sun
- abstract@[open-review](#): Multi-agent reinforcement learning has drawn increasing attention in practice, e.g., robotics and automatic driving, as it can explore optimal policies using samples generated by interacting with the environment. However, high reward uncertainty still remains a problem when we want to train a satisfactory model, because obtaining high-quality reward feedback is usually expensive and even infeasible. To handle this issue, previous methods mainly focus on passive reward correction. At the same time, recent active reward estimation methods have proven to be a recipe for reducing the effect of reward uncertainty. In this paper, we propose a novel Distributional Reward Estimation framework for effective Multi-Agent Reinforcement Learning (DRE-MARL). Our main idea is to design the multi-action-branch reward estimation and policy-weighted reward aggregation for stabilized training. Specifically, we design the multi-action-branch reward estimation to model reward distributions on all action branches. Then we utilize reward aggregation to obtain stable updating signals during training. Our intuition is that consideration of all possible consequences of actions could be useful for learning policies. The superiority of the DRE-MARL is demonstrated using benchmark multi-agent scenarios, compared with the SOTA baselines in terms of both effectiveness and robustness.

## [Unsupervised Cross-Domain Imitation Learning](#)

- Tim Franzmeyer · Philip Torr · JoÃ±o Henriques
- abstract@[open-review](#): We study how an autonomous agent learns to perform a task from demonstrations in a different domain, such as a different environment or different agent. Such cross-domain imitation learning is required to, for example, train an artificial agent from demonstrations of a human expert. We propose a scalable framework that enables cross-domain imitation learning without access to additional demonstrations or further domain knowledge. We jointly train the learner agent's policy and learn a mapping between the learner and expert domains with adversarial training. We effect this by using a mutual information criterion to find an embedding of the expert's state space that contains task-relevant information and is invariant to domain

specifics. This step significantly simplifies estimating the mapping between the learner and expert domains and hence facilitates end-to-end learning. We demonstrate successful transfer of policies between considerably different domains, without extra supervision such as additional demonstrations, and in situations where other methods fail.

## [Convergence beyond the over-parameterized regime using Rayleigh quotients](#)

- David Robin · Kevin Scaman · marc lelarge
- abstract@[open-review](#): In this paper, we present a new strategy to prove the convergence of Deep Learning architectures to a zero training (or even testing) loss by gradient flow. Our analysis is centered on the notion of Rayleigh quotients in order to prove Kurdyka-Łojasiewicz inequalities for a broader set of neural network architectures and loss functions. We show that Rayleigh quotients provide a unified view for several convergence analysis techniques in the literature. Our strategy produces a proof of convergence for various examples of parametric learning. In particular, our analysis does not require the number of parameters to tend to infinity, nor the number of samples to be finite, thus extending to test loss minimization and beyond the over-parameterized regime.

## [Function Classes for Identifiable Nonlinear Independent Component Analysis](#)

- Simon Buchholz · Michel Besserve · Bernhard Schölkopf
- abstract@[open-review](#): Unsupervised learning of latent variable models (LVMs) is widely used to represent data in machine learning. When such model reflects the ground truth factors and the mechanisms mapping them to observations, there is reason to expect that such models allow generalisation in downstream tasks. It is however well known that such identifiability guarantees are typically not achievable without putting constraints on the model class. This is notably the case for nonlinear Independent Component Analysis, in which the LVM maps statistically independent variables to observations via a deterministic nonlinear function. Several families of spurious solutions fitting perfectly the data, but that do not correspond to the ground truth factors can be constructed in generic settings. However, recent work suggests that constraining the function class of such models may promote identifiability. Specifically, function classes with constraints on their partial derivatives, gathered in the Jacobian matrix, have been proposed, such as orthogonal coordinate transformations (OCT), which impose orthogonality of the Jacobian columns. In the present work, we prove that a subclass of these transformations, conformal maps, is identifiable and provide novel theoretical results suggesting that OCTs have properties that prevent families of spurious solutions to spoil identifiability in a generic setting.

## [Tree ensemble kernels for Bayesian optimization with known constraints over mixed-feature spaces](#)

- Alexander Thebelt · Calvin Tsay · Robert Lee · Nathan Sudermann-Merx · David Walz · Behrang Shafei · Ruth Misener
- abstract@[open-review](#): Tree ensembles can be well-suited for black-box optimization tasks such as algorithm tuning and neural architecture search, as they achieve good predictive performance with little or no manual tuning, naturally handle discrete feature spaces, and are relatively insensitive to outliers in the training data. Two well-known challenges in using tree ensembles for black-box optimization are (i) effectively quantifying model uncertainty for exploration and (ii) optimizing over the piece-wise constant acquisition function. To address both points simultaneously, we propose using the kernel interpretation of tree ensembles as a Gaussian Process prior to obtain model variance estimates, and we develop a compatible optimization formulation for the acquisition function. The latter further allows us to seamlessly integrate known constraints to improve sampling efficiency by considering domain-knowledge in engineering settings and modeling search space symmetries, e.g. hierarchical relationships in neural architecture search. Our framework performs as well as state-of-the-art methods for unconstrained black-box optimization over continuous/discrete features and outperforms competing methods for problems combining mixed-variable feature spaces and known input constraints.

## [OnePose++: Keypoint-Free One-Shot Object Pose Estimation without CAD Models](#)

- Xingyi He · Jiaming Sun · Yuang Wang · Di Huang · Hujun Bao · Xiaowei Zhou
- abstract@[open-review](#): We propose a new method for object pose estimation without CAD models. The previous feature-matching-based method OnePose has shown promising results under a one-shot setting, eliminating the object-specific training and aiming for the pose estimation of arbitrary unseen objects. However, OnePose relies on detecting repeatable image keypoints and is thus prone to fail on low-textured objects. We propose to remove the need for keypoint detection with a keypoint-free SfM and pose estimation pipeline. Built upon the keypoint-free feature matching method LoFTR, our novel keypoint-free SfM reconstructs object semi-dense point clouds, which provides complete 3D models for pose estimation of low-textured objects. During pose estimation, our novel keypoint-free 2D-3D matching network matches the reconstructed point cloud with the query image to build 2D-3D correspondences for object pose estimation. The keypoint-free SfM and pose estimation pipeline enables pose estimation of low-textured objects under the one-shot setting, making it more applicable in the real world. We demonstrate that the proposed method outperforms existing one-shot CAD-model-free methods by a large margin and even achieves comparable results with CAD-model-based methods on LINEMOD. We also collect a new dataset composed of 80 sequences of 40 low-textured objects to boost future research on one-shot object pose estimation. Code and data will be made available upon the publication of this paper.

## [When to Trust Your Simulator: Dynamics-Aware Hybrid Offline-and-Online Reinforcement Learning](#)

- Haoyi Niu · shubham sharma · Yiwen Qiu · Ming Li · Guyue Zhou · Jianming HU · Xianyuan Zhan
- abstract@[open-review](#): Learning effective reinforcement learning (RL) policies to solve real-world complex tasks can be quite challenging without a high-fidelity simulation environment. In most cases, we are only given imperfect simulators with simplified dynamics, which inevitably lead to severe sim-to-real gaps in RL policy learning. The recently emerged field of offline RL provides another possibility to learn policies directly from pre-collected historical data. However, to achieve reasonable performance, existing offline RL algorithms need impractically large offline data with sufficient state-action space coverage for training. This brings up a new question: is it possible to combine learning from limited real data in offline RL and unrestricted exploration through imperfect simulators in online RL to address the drawbacks of both approaches? In this study, we propose the Dynamics-Aware Hybrid Offline-and-Online Reinforcement Learning (H2O) framework to provide an affirmative answer to this question. H2O introduces a dynamics-aware policy evaluation scheme, which adaptively penalizes the Q function learning on simulated state-action pairs with large dynamics gaps, while also simultaneously allowing learning from a fixed real-world dataset. Through extensive simulation and real-world tasks, as well as theoretical analysis, we demonstrate the superior performance of H2O against other cross-domain online and offline RL algorithms. H2O provides a brand new hybrid offline-and-online RL paradigm, which can potentially shed light on future RL algorithm design for solving practical real-world tasks.

## [Prototypical VoteNet for Few-Shot 3D Point Cloud Object Detection](#)

- Shizhen Zhao · Xiaojuan Qi
- abstract@[open-review](#): Most existing 3D point cloud object detection approaches heavily rely on large amounts of labeled training data. However, the labeling process is costly and time-consuming. This paper considers few-shot 3D point cloud object detection, where only a few annotated samples of novel classes are needed with abundant samples of base classes. To this end, we propose Prototypical VoteNet to recognize and localize novel instances, which incorporates two new modules: Prototypical Vote Module (PVM) and Prototypical Head Module (PHM). Specifically, as the 3D basic geometric structures can be shared among categories, PVM is designed to leverage class-agnostic geometric prototypes, which are learned from base classes, to refine local features of novel categories. Then PHM is proposed to utilize class prototypes to enhance the global feature of each object, facilitating subsequent object localization and classification, which is trained by the episodic training strategy. To evaluate the model in this new setting, we contribute two new benchmark datasets, FS-ScanNet and FS-SUNRGBD. We conduct extensive experiments to demonstrate the effectiveness of Prototypical

VoteNet, and our proposed method shows significant and consistent improvements compared to baselines on two benchmark datasets. Our code, as well as the new benchmark, will be released to facilitate future works.

## [Collaborative Linear Bandits with Adversarial Agents: Near-Optimal Regret Bounds](#)

- Aritra Mitra · Arman Adibi · George J. Pappas · Hamed Hassani
- abstract@[open-review](#): We consider a linear stochastic bandit problem involving  $M$  agents that can collaborate via a central server to minimize regret. A fraction  $\alpha$  of these agents are adversarial and can act arbitrarily, leading to the following tension: while collaboration can potentially reduce regret, it can also disrupt the process of learning due to adversaries. In this work, we provide a fundamental understanding of this tension by designing new algorithms that balance the exploration-exploitation trade-off via carefully constructed robust confidence intervals. We also complement our algorithms with tight analyses. First, we develop a robust collaborative phased elimination algorithm that achieves  $\tilde{O}(\alpha + 1/\sqrt{M})\sqrt{dT}$  regret for each good agent; here,  $d$  is the model-dimension and  $T$  is the horizon. For small  $\alpha$ , our result thus reveals a clear benefit of collaboration despite adversaries. Using an information-theoretic argument, we then prove a matching lower bound, thereby providing the first set of tight, near-optimal regret bounds for collaborative linear bandits with adversaries. Furthermore, by leveraging recent advances in high-dimensional robust statistics, we significantly extend our algorithmic ideas and results to (i) the generalized linear bandit model that allows for non-linear observation maps; and (ii) the contextual bandit setting that allows for time-varying feature vectors.

## [Contextual Bandits with Knapsacks for a Conversion Model](#)

- Zhen LI · Gilles Stoltz
- abstract@[open-review](#): We consider contextual bandits with knapsacks, with an underlying structure between rewards generated and cost vectors suffered. We do so motivated by sales with commercial discounts. At each round, given the stochastic i.i.d. context  $\mathbf{x}_t$  and the arm picked  $a_t$  (corresponding, e.g., to a discount level), a customer conversion may be obtained, in which case a reward  $r(a, \mathbf{x}_t)$  is gained and vector costs  $\mathbf{c}(a_t, \mathbf{x}_t)$  are suffered (corresponding, e.g., to losses of earnings). Otherwise, in the absence of a conversion, the reward and costs are null. The reward and costs achieved are thus coupled through the binary variable measuring conversion or the absence thereof. This underlying structure between rewards and costs is different from the linear structures considered by Agrawal and Devanur [2016] (but we show that the techniques introduced in the present article may also be applied to the case of these linear structures). The adaptive policies exhibited in this article solve at each round a linear program based on upper-confidence estimates of the probabilities of conversion given  $a$  and  $\mathbf{x}$ . This kind of policy is most natural and achieves a regret bound of the typical order  $(\mathrm{OPT}/B) \sqrt{T}$ , where  $B$  is the total budget allowed,  $\mathrm{OPT}$  is the optimal expected reward achievable by a static policy, and  $T$  is the number of rounds.

## [Factuality Enhanced Language Models for Open-Ended Text Generation](#)

- Nayeon Lee · Wei Ping · Peng Xu · Mostafa Patwary · Mohammad Shoeybi · Bryan Catanzaro
- abstract@[open-review](#): Pretrained language models~(LMs) can easily generate text with nonfactual information. In this work, we measure and improve the factual accuracy of large-scale LMs for open-ended text generation. We design the FactualPrompts test set and metrics to measure the factuality of LM generations. Based on that, we study the factual accuracy of LMs with parameter sizes ranging from 126M to 530B. Interestingly, we find that larger LMs are more factual than smaller ones, although a previous study suggests that larger LMs can be less truthful in terms of misconceptions. In addition, popular sampling algorithms~(e.g., top-p) in open-ended text generation can reduce the factuality due to the "uniform randomness" introduced at every sampling step. We propose a factual-nucleus sampling algorithm that dynamically adapts the randomness to improve the factuality of generation while maintaining quality. Furthermore, we analyze the inefficiencies of the standard training method in learning correct associations between entities from factual text corpus~(e.g., Wikipedia). We propose a factuality-enhanced training method that uses TopicPrefix for better awareness of facts and sentence completion as the training objective, which vastly reduces the factual errors from LMs.

## [A Boosting Approach to Reinforcement Learning](#)

- Nataly Brukhim · Elad Hazan · Karan Singh
- abstract@[open-review](#): Reducing reinforcement learning to supervised learning is a well-studied and effective approach that leverages the benefits of compact function approximation to deal with large-scale Markov decision processes. Independently, the boosting methodology (e.g. AdaBoost) has proven to be indispensable in designing efficient and accurate classification algorithms by combining rough and inaccurate rules-of-thumb. In this paper, we take a further step: we reduce reinforcement learning to a sequence of weak learning problems. Since weak learners perform only marginally better than random guesses, such subroutines constitute a weaker assumption than the availability of an accurate supervised learning oracle. We prove that the sample complexity and running time bounds of the proposed method do not explicitly depend on the number of states. While existing results on boosting operate on convex losses, the value function over policies is non-convex. We show how to use a non-convex variant of the Frank-Wolfe method for boosting, that additionally improves upon the known sample complexity and running time bounds even for reductions to supervised learning.

## [Decomposed Knowledge Distillation for Class-incremental Semantic Segmentation](#)

- Donghyeon Baek · Youngmin Oh · Sanghoon Lee · Junghyun Lee · Bumsub Ham
- abstract@[open-review](#): Class-incremental semantic segmentation (CISS) labels each pixel of an image with a corresponding object/stuff class continually. To this end, it is crucial to learn novel classes incrementally without forgetting previously learned knowledge. Current CISS methods typically use a knowledge distillation (KD) technique for preserving classifier logits, or freeze a feature extractor, to avoid the forgetting problem. The strong constraints, however, prevent learning discriminative features for novel classes. We introduce a CISS framework that alleviates the forgetting problem and facilitates learning novel classes effectively. We have found that a logit can be decomposed into two terms. They quantify how likely an input belongs to a particular class or not, providing a clue for a reasoning process of a model. The KD technique, in this context, preserves the sum of two terms ( $i.e.$ , a class logit), suggesting that each could be changed and thus the KD does not imitate the reasoning process. To impose constraints on each term explicitly, we propose a new decomposed knowledge distillation (DKD) technique, improving the rigidity of a model and addressing the forgetting problem more effectively. We also introduce a novel initialization method to train new classifiers for novel classes. In CISS, the number of negative training samples for novel classes is not sufficient to discriminate old classes. To mitigate this, we propose to transfer knowledge of negatives to the classifiers successively using an auxiliary classifier, boosting the performance significantly. Experimental results on standard CISS benchmarks demonstrate the effectiveness of our framework.

## [Learning Generalizable Models for Vehicle Routing Problems via Knowledge Distillation](#)

- Jieyi Bi · Yining Ma · Jiahai Wang · Zhiguang Cao · Jinbiao Chen · Yuan Sun · Yeow Meng Chee
- abstract@[open-review](#): Recent neural methods for vehicle routing problems always train and test the deep models on the same instance distribution (i.e., uniform). To tackle the consequent cross-distribution generalization concerns, we bring the knowledge distillation to this field and propose an Adaptive Multi-Distribution Knowledge Distillation (AMDKD) scheme for learning more generalizable deep models. Particularly, our AMDKD leverages various knowledge from multiple teachers trained on exemplar distributions to yield a light-weight yet generalist student model. Meanwhile, we equip AMDKD with an adaptive strategy that allows the student to concentrate on difficult distributions, so as to absorb hard-to-master knowledge more effectively. Extensive experimental results show that, compared with the baseline neural methods, our AMDKD is able to achieve competitive results on both unseen

in-distribution and out-of-distribution instances, which are either randomly synthesized or adopted from benchmark datasets (i.e., TSPLIB and CVRPLIB). Notably, our AMDKD is generic, and consumes less computational resources for inference.

## [A Regret-Variance Trade-Off in Online Learning](#)

- Dirk van der Hoeven · Nikita Zhivotovskiy · Nicolás Cesa-Bianchi
- abstract@[open-review](#): We consider prediction with expert advice for strongly convex and bounded losses, and investigate trade-offs between regret and "variance" (i.e., squared difference of learner's predictions and best expert predictions). With  $K$  experts, the Exponentially Weighted Average (EWA) algorithm is known to achieve  $O(\log K)$  regret. We prove that a variant of EWA either achieves a  $\text{negative}$  regret (i.e., the algorithm outperforms the best expert), or guarantees a  $O(\log K)$  bound on  $\text{both}$  variance and regret. Building on this result, we show several examples of how variance of predictions can be exploited in learning. In the online to batch analysis, we show that a large empirical variance allows to stop the online to batch conversion early and outperform the risk of the best predictor in the class. We also recover the optimal rate of model selection aggregation when we do not consider early stopping. In online prediction with corrupted losses, we show that the effect of corruption on the regret can be compensated by a large variance. In online selective sampling, we design an algorithm that samples less when the variance is large, while guaranteeing the optimal regret bound in expectation. In online learning with abstention, we use a similar term as the variance to derive the first high-probability  $O(\log K)$  regret bound in this setting. Finally, we extend our results to the setting of online linear regression.

## [Learning on the Edge: Online Learning with Stochastic Feedback Graphs](#)

- Emmanuel Esposito · Federico Fusco · Dirk van der Hoeven · Nicolás Cesa-Bianchi
- abstract@[open-review](#): The framework of feedback graphs is a generalization of sequential decision-making with bandit or full information feedback. In this work, we study an extension where the directed feedback graph is stochastic, following a distribution similar to the classical Erdős-Rényi model. Specifically, in each round every edge in the graph is either realized or not with a distinct probability for each edge. We prove regret bounds of order  $\min\{\min_{\{w\}} \sqrt{(\alpha_w w)^T}, \min_{\{w\}} (\delta_w w)^{1/3} T^{2/3}\}$  (ignoring logarithmic factors), where  $\alpha_w$  and  $\delta_w$  are graph-theoretic quantities measured on the support of the stochastic feedback graph  $G$  with edge probabilities thresholded at  $w$ . Our result, which holds without any preliminary knowledge about  $G$ , requires the learner to observe only the realized out-neighborhood of the chosen action. When the learner is allowed to observe the realization of the entire graph (but only the losses in the out-neighborhood of the chosen action), we derive a more efficient algorithm with a tighter bound featuring a dependence on weighted versions of the independence and weak domination numbers.

## [NeuPhysics: Editable Neural Geometry and Physics from Monocular Videos](#)

- Yi-Ling Qiao · Alexander Gao · Ming Lin
- abstract@[open-review](#): We present a method for learning geometry and physics parameters of a dynamic scene requiring only a monocular RGB video. Our approach uses a hybrid representation of neural fields and hexahedra mesh, enabling objects in the scene to be interactively edited, and synthesized from novel views. To decouple the learning of underlying scene geometry from dynamic motion, we learn a time-invariant signed distance function which serves as a reference frame, as well as an associated deformation field that is conditioned on each time step. We design a two-way conversion between the neural field and corresponding mesh representation, which allows us to bridge the neural representation with a differentiable physics simulator, and therefore estimate physics parameters from the source video, by minimizing a cycle consistency loss. This flexible, hybrid representation also allows a user to easily edit 3D objects from the source video by directly editing the recovered hexahedra mesh, and propagating this operation back to the neural field. In Experiments, our method achieves higher-quality mesh and video reconstruction of dynamic scenes compared to other competitive Neural Field methods. Finally, we provide extensive examples which demonstrate our method's ability to extract useful 3D representations of dynamic scenes from videos captured with consumer-grade cameras.

## [Lethal Dose Conjecture on Data Poisoning](#)

- Wenxiao Wang · Alexander Levine · Soheil Feizi
- abstract@[open-review](#): Data poisoning considers an adversary that distorts the training set of machine learning algorithms for malicious purposes. In this work, we bring to light one conjecture regarding the fundamentals of data poisoning, which we call the Lethal Dose Conjecture. The conjecture states: If  $n$  clean training samples are needed for accurate predictions, then in a size- $N$  training set, only  $\Theta(N/n)$  poisoned samples can be tolerated while ensuring accuracy. Theoretically, we verify this conjecture in multiple cases. We also offer a more general perspective of this conjecture through distribution discrimination. Deep Partition Aggregation (DPA) and its extension, Finite Aggregation (FA) are recent approaches for provable defenses against data poisoning, where they predict through the majority vote of many base models trained from different subsets of training set using a given learner. The conjecture implies that both DPA and FA are (asymptotically) optimal---if we have the most data-efficient learner, they can turn it into one of the most robust defenses against data poisoning. This outlines a practical approach to developing stronger defenses against poisoning via finding data-efficient learners. Empirically, as a proof of concept, we show that by simply using different data augmentations for base learners, we can respectively double and triple the certified robustness of DPA on CIFAR-10 and GTSRB without sacrificing accuracy.

## [Disentangling Causal Effects from Sets of Interventions in the Presence of Unobserved Confounders](#)

- Olivier Jeunen · Ciarán Gilligan-Lee · Rishabh Mehrotra · Mounia Lalmas
- abstract@[open-review](#): The ability to answer causal questions is crucial in many domains, as causal inference allows one to understand the impact of interventions. In many applications, only a single intervention is possible at a given time. However, in some important areas, multiple interventions are concurrently applied. Disentangling the effects of single interventions from jointly applied interventions is a challenging task---especially as simultaneously applied interventions can interact. This problem is made harder still by unobserved confounders, which influence both treatments and outcome. We address this challenge by aiming to learn the effect of a single-intervention from both observational data and sets of interventions. We prove that this is not generally possible, but provide identification proofs demonstrating that it can be achieved under non-linear continuous structural causal models with additive, multivariate Gaussian noise---even when unobserved confounders are present. Importantly, we show how to incorporate observed covariates and learn heterogeneous treatment effects. Based on the identifiability proofs, we provide an algorithm that learns the causal model parameters by pooling data from different regimes and jointly maximising the combined likelihood. The effectiveness of our method is empirically demonstrated on both synthetic and real-world data.

## [Mask Matching Transformer for Few-Shot Segmentation](#)

- siyu jiao · Gengwei Zhang · Shant Navasardyan · Ling Chen · Yao Zhao · Yunchao Wei · Honghui Shi
- abstract@[open-review](#): In this paper, we aim to tackle the challenging few-shot segmentation task from a new perspective. Typical methods follow the paradigm to firstly learn prototypical features from support images and then match query features in pixel-level to obtain segmentation results. However, to obtain satisfactory segments, such a paradigm needs to couple the learning of the matching operations with heavy segmentation modules, limiting the flexibility of design and increasing the learning complexity. To alleviate this issue, we propose Mask Matching Transformer (MM-Former), a new paradigm for the few-shot segmentation task. Specifically, MM-Former first uses a class-agnostic segmenter to decompose the query image into multiple segment proposals. Then, a simple matching mechanism is applied to merge the related segment proposals into the final mask guided by the support images. The advantages of our MM-Former are two-fold. First, the MM-Former follows the paradigm of decomposing first and then blending, allowing

our method to benefit from the advanced potential objects segmenter to produce high-quality mask proposals for query images. Second, the mission of prototypical features is relaxed to learn coefficients to fuse correct ones within a proposal pool, making the MM-Former well generalized to complex scenarios or cases. We conduct extensive experiments on the popular COCO-\$20^{\text{th}}\$ and Pascal-\$5^{\text{th}}\$ benchmarks. Competitive results well demonstrate the effectiveness and the generalization ability of our MM-Former. Code is available in the supplementary materials.

## [Test-Time Training with Masked Autoencoders](#)

- Yossi Gandelsman · Yu Sun · Xinlei Chen · Alexei Efros
- abstract@[open-review](#): Prior work has shown masked autoencoding as an effective self-supervised task across many visual distributions with sufficient training data. We use masked autoencoding to train on each unlabeled test sample as it arrives at test time, before making a prediction. This simple method improves generalization of predictive models on many visual benchmarks of unknown distributions, where input images are not covered by any training data. Our theoretical analysis explains how test-time training with autoencoding helps a linear model under distribution shifts.

## [Image Inpainting models are Few-Shot Learners \(Given the Right Data\)](#)

- Amir Bar · Yossi Gandelsman · Trevor Darrell · Amir Globerson · Alexei Efros
- abstract@[open-review](#): How does one adapt a pre-trained visual model to novel downstream tasks without task-specific finetuning or any model modification? Taking inspiration from prompting in NLP systems, this paper investigates Visual Prompting: given input-output image example(s) of a new task at test time and a new input image, the goal is to automatically produce the correct output image, consistent with the proposed task. We show that posing this problem as a simple image inpainting task - literally just filling in a hole in a concatenated visual prompt image - turns out to be surprisingly effective, given that the inpainting algorithm has been trained on the right data. We train masked auto-encoding models on a new dataset that we curated - 88k unlabeled figures from academic papers sources on Arxiv. We apply visual prompting to these pretrained models and demonstrate results on various downstream tasks, including foreground segmentation, single object detection, colorization, edge detection, etc. All our code, models, and dataset will be made available.

## [Reproducibility in Optimization: Theoretical Framework and Limits](#)

- Kwangjun Ahn · Prateek Jain · Ziwei Ji · Satyen Kale · Praneeth Netrapalli · Gil I Shamir
- abstract@[open-review](#): We initiate a formal study of reproducibility in optimization. We define a quantitative measure of reproducibility of optimization procedures in the face of noisy or error-prone operations such as inexact or stochastic gradient computations or inexact initialization. We then analyze several convex optimization settings of interest such as smooth, non-smooth, and strongly-convex objective functions and establish tight bounds on the limits of reproducibility in each setting. Our analysis reveals a fundamental trade-off between computation and reproducibility: more computation is necessary (and sufficient) for better reproducibility.

## [AUTOMATA: Gradient Based Data Subset Selection for Compute-Efficient Hyper-parameter Tuning](#)

- Krishnateja Killamsetty · Guttu Sai Abhishek · Aakriti Lnu · Alexandre Evfimievski · Lucian Popa · Ganesh Ramakrishnan · Rishabh Iyer
- abstract@[open-review](#): Deep neural networks have seen great success in recent years; however, training a deep model is often challenging as its performance heavily depends on the hyper-parameters used. In addition, finding the optimal hyper-parameter configuration, even with state-of-the-art (SOTA) hyper-parameter optimization (HPO) algorithms, can be time-consuming, requiring multiple training runs over the entire dataset for different possible sets of hyper-parameters. Our central insight is that using an informative subset of the dataset for model training runs involved in hyper-parameter optimization, allows us to find the optimal hyper-parameter configuration significantly faster. In this work, we propose AUTOMATA, a gradient-based subset selection framework for hyper-parameter tuning. We empirically evaluate the effectiveness of AUTOMATA in hyper-parameter tuning through several experiments on real-world datasets in the text, vision, and tabular domains. Our experiments show that using gradient-based data subsets for hyper-parameter tuning achieves significantly faster turnaround times and speedups of 3\$\times\$-30\$\times\$ while achieving comparable performance to the hyper-parameters found using the entire dataset.

## [Orient: Submodular Mutual Information Measures for Data Subset Selection under Distribution Shift](#)

- Athresh Karanam · Krishnateja Killamsetty · Harsha Kokel · Rishabh Iyer
- abstract@[open-review](#): Real-world machine-learning applications require robust models that generalize well to distribution shift settings, which is typical of real-world situations. Domain adaptation techniques aim to address this issue of distribution shift by minimizing the disparities between domains to ensure that the model trained on the source domain performs well on the target domain. Nevertheless, the existing domain adaptation methods are computationally very expensive. In this work, we aim to improve the efficiency of existing supervised domain adaptation (SDA) methods by using a subset of source data that is similar to target data for faster model training. Specifically, we propose ORIENT, a subset selection framework that uses the submodular mutual information (SMI) functions to select a source data subset similar to the target data for faster training. Additionally, we demonstrate how existing robust subset selection strategies, such as GLISTER, GRADMATCH, and CRAIG, when used with a held-out query set, fit within our proposed framework and demonstrate the connections with them. Finally, we empirically demonstrate that SDA approaches like d-SNE, CCSA, and standard Cross-entropy training, when employed together with ORIENT, achieve a) faster training and b) better performance on the target data.

## [Enhanced Latent Space Blind Model for Real Image Denoising via Alternative Optimization](#)

- Chao Ren · Yizhong Pan · Jie Huang
- abstract@[open-review](#): We propose a novel enhanced latent space blind model based deep unfolding network, namely ScaedNet, for complex real image denoising. Our approach is derived by introducing latent space, noise information, and guidance constraint into the denoising cost function. A self-correction alternative optimization algorithm is proposed to split the novel cost function into three alternative sub-problems, i.e., guidance representation (GR), degradation estimation (DE) and reconstruction (RE) subproblems. Finally, we implement the optimization process by a deep unfolding network consisting of GR, DE and RE networks. For higher performance of the DE network, a novel parameter-free noise feature adaptive enhancement (NFAE) layer is proposed. To synchronously and dynamically realize internal-external feature information mining in the RE network, a novel feature multi-modulation attention (FM2A) module is proposed. Our approach thereby leverages the advantages of deep learning, while also benefiting from the principled denoising provided by the classical model-based formulation. To the best of our knowledge, our enhanced latent space blind model, optimization scheme, NFAE and FM2A have not been reported in the previous literature. Experimental results show the promising performance of ScaedNet on real image denoising.

## [Unsupervised Multi-View Object Segmentation Using Radiance Field Propagation](#)

- Xinhang Liu · Jiaben Chen · Huai Yu · Yu-Wing Tai · Chi-Keung Tang
- abstract@[open-review](#): We present radiance field propagation (RFP), a novel approach to segmenting objects in 3D during reconstruction given only unlabeled multi-view images of a scene. RFP is derived from emerging neural radiance field-based techniques, which jointly encodes semantics with appearance and geometry. The core of our method is a novel propagation strategy for individual objects' radiance fields with a bidirectional photometric loss, enabling an unsupervised partitioning of a scene into salient or meaningful regions corresponding to different object instances. To better handle

complex scenes with multiple objects and occlusions, we further propose an iterative expectation-maximization algorithm to refine object masks. To the best of our knowledge, RFP is the first unsupervised approach for tackling 3D scene object segmentation for neural radiance field (NeRF) without any supervision, annotations, or other cues such as 3D bounding boxes and prior knowledge of object class. Experiments demonstrate that RFP achieves feasible segmentation results that are more accurate than previous unsupervised image/scene segmentation approaches, and are comparable to existing supervised NeRF-based methods. The segmented object representations enable individual 3D object editing operations. Codes and datasets will be made publicly available.

## [One Model to Edit Them All: Free-Form Text-Driven Image Manipulation with Semantic Modulations](#)

- Yiming Zhu · Hongyu Liu · Yibing Song · Ziyang Yuan · Xintong Han · Chun Yuan · Qifeng Chen · Jue Wang
- abstract@[open-review](#): Free-form text prompts allow users to describe their intentions during image manipulation conveniently. Based on the visual latent space of StyleGAN[19]and text embedding space of CLIP[28], studies focus on how to map these two latent spaces for text-driven attribute manipulations. Currently, the latent mapping between these two spaces is empirically designed and confines that each manipulation model can only tackle one fixed text prompt. In this paper, we propose a method named Free-Form CLIP (FFCLIP), aiming to establish an automatic latent mapping so that one manipulation model handles free-form text prompts. Our FFCLIP has a cross-modality semantic modulation module containing semantic alignment and injection. The semantic alignment performs the automatic latent mapping via linear transformations with a cross attention mechanism. After alignment, we inject semantics from text prompt embeddings to the StyleGAN latent space. For one type of image (e.g., human portrait'), one FFCLIP model can be learned with free-form text prompts. Meanwhile, we observe that although each training text prompt only contains a single semantic meaning, FFCLIP can leverage text prompts with multiple semantic meanings for image manipulation. In the experiments, we evaluate FFCLIP on three types of images (i.e.,human portraits', cars', andchurches'). Both visual and numerical results show that FFCLIP effectively produces semantically accurate and visually realistic images.

## [Preservation of the Global Knowledge by Not-True Distillation in Federated Learning](#)

- Gihun Lee · Minchan Jeong · Yongjin Shin · Sangmin Bae · Se-Young Yun
- abstract@[open-review](#): In federated learning, a strong global model is collaboratively learned by aggregating clients' locally trained models. Although this precludes the need to access clients' data directly, the global model's convergence often suffers from data heterogeneity. This study starts from an analogy to continual learning and suggests that forgetting could be the bottleneck of federated learning. We observe that the global model forgets the knowledge from previous rounds, and the local training induces forgetting the knowledge outside of the local distribution. Based on our findings, we hypothesize that tackling down forgetting will relieve the data heterogeneity problem. To this end, we propose a novel and effective algorithm, Federated Not-True Distillation (FedNTD), which preserves the global perspective on locally available data only for the not-true classes. In the experiments, FedNTD shows state-of-the-art performance on various setups without compromising data privacy or incurring additional communication costs.

## [Towards Hard-pose Virtual Try-on via 3D-aware Global Correspondence Learning](#)

- Zaiyu Huang · Hanhui Li · Zhenyu Xie · Michael Kampffmeyer · qingling Cai · Xiaodan Liang
- abstract@[open-review](#): In this paper, we target image-based person-to-person virtual try-on in the presence of diverse poses and large viewpoint variations. Existing methods are restricted in this setting as they estimate garment warping flows mainly based on 2D poses and appearance, which omits the geometric prior of the 3D human body shape. Moreover, current garment warping methods are confined to localized regions, which makes them ineffective in capturing long-range dependencies and results in inferior flows with artifacts. To tackle these issues, we present 3D-aware global correspondences, which are reliable flows that jointly encode global semantic correlations, local deformations, and geometric priors of 3D human bodies. Particularly, given an image pair depicting the source and target person, (a) we first obtain their pose-aware and high-level representations via two encoders, and introduce a coarse-to-fine decoder with multiple refinement modules to predict the pixel-wise global correspondence. (b) 3D parametric human models inferred from images are incorporated as priors to regularize the correspondence refinement process so that our flows can be 3D-aware and better handle variations of pose and viewpoint. (c) Finally, an adversarial generator takes the garment warped by the 3D-aware flow, and the image of the target person as inputs, to synthesize the photo-realistic try-on result. Extensive experiments on public benchmarks and our selected HardPose test set demonstrate the superiority of our method against state-of-the-art try-on approaches.

## [Distilling Representations from GAN Generator via Squeeze and Span](#)

- Yu Yang · Xiaotian Cheng · Chang Liu · Hakan Bilen · Xiangyang Ji
- abstract@[open-review](#): In recent years, generative adversarial networks (GANs) have been an actively studied topic and shown to successfully produce high-quality realistic images in various domains. The controllable synthesis ability of GAN generators suggests that they maintain informative, disentangled, and explainable image representations, but leveraging and transferring their representations to downstream tasks is largely unexplored. In this paper, we propose to distill knowledge from GAN generators by squeezing and spanning their representations. We \text{squeeze} the generator features into representations that are invariant to semantic-preserving transformations through a network before they are distilled into the student network. We \text{span} the distilled representation of the synthetic domain to the real domain by also using real training data to remedy the mode collapse of GANs and boost the student network performance in a real domain. Experiments justify the efficacy of our method and reveal its great significance in self-supervised representation learning. Code will be made public.

## [Selective compression learning of latent representations for variable-rate image compression](#)

- Jooyoung Lee · Seyoon Jeong · Munchurl Kim
- abstract@[open-review](#): Recently, many neural network-based image compression methods have shown promising results superior to the existing tool-based conventional codecs. However, most of them are often trained as separate models for different target bit rates, thus increasing the model complexity. Therefore, several studies have been conducted for learned compression that supports variable rates with single models, but they require additional network modules, layers, or inputs that often lead to complexity overhead, or do not provide sufficient coding efficiency. In this paper, we firstly propose a selective compression method that partially encodes the latent representations in a fully generalized manner for deep learning-based variable-rate image compression. The proposed method adaptively determines essential representation elements for compression of different target quality levels. For this, we first generate a 3D importance map as the nature of input content to represent the underlying importance of the representation elements. The 3D importance map is then adjusted for different target quality levels using importance adjustment curves. The adjusted 3D importance map is finally converted into a 3D binary mask to determine the essential representation elements for compression. The proposed method can be easily integrated with the existing compression models with a negligible amount of overhead increase. Our method can also enable continuously variable-rate compression via simple interpolation of the importance adjustment curves among different quality levels. The extensive experimental results show that the proposed method can achieve comparable compression efficiency as those of the separately trained reference compression models and can reduce decoding time owing to the selective compression.

## [Model-Based Offline Reinforcement Learning with Pessimism-Modulated Dynamics Belief](#)

- Kaiyang Guo · Shao Yunfeng · Yanhui Geng
- abstract@[open-review](#): Model-based offline reinforcement learning (RL) aims to find highly rewarding policy, by leveraging a previously collected static dataset and a learned dynamics model. While the dynamics model is learned by reusing the static dataset, its generalization ability hopefully promotes

policy learning if properly utilized. To that end, several works propose to quantify the uncertainty of predicted dynamics, and explicitly apply it to penalize reward. However, as the dynamics has different implication than the reward, characterizing the impact of dynamics uncertainty through reward penalty may incur unexpected tradeoff between model utilization and risk avoidance. In this work, we instead maintain a belief distribution over dynamics, and evaluate/optimize policy through biased sampling from the belief. The sampling procedure, biased towards pessimism, is derived based on an alternating Markov game formulation of offline RL. We formally show that the biased sampling naturally induces an updated dynamics belief with policy-dependent reweighting factor, termed Pessimism-Modulated Dynamics Belief. To improve policy, we devise an iterative regularized policy optimization algorithm for the game, with guarantee of monotonous improvement under certain condition. To make practical, we further devise an offline RL algorithm to approximately find the solution. Empirical results show that the proposed approach achieves state-of-the-art performance on a wide range of offline RL benchmark tasks.

## [Efficient identification of informative features in simulation-based inference](#)

- Jonas Beck · Michael Deistler · Yves Bernaerts · Jakob H Macke · Philipp Berens
- abstract@[open-review](#): Simulation-based Bayesian inference (SBI) can be used to estimate the parameters of complex mechanistic models given observed model outputs without requiring access to explicit likelihood evaluations. A prime example for the application of SBI in neuroscience involves estimating the parameters governing the response dynamics of Hodgkin-Huxley (HH) models from electrophysiological measurements, by inferring a posterior over the parameters that is consistent with a set of observations. To this end, many SBI methods employ a set of summary statistics or scientifically interpretable features to estimate a surrogate likelihood or posterior. However, currently, there is no way to identify how much each summary statistic or feature contributes to reducing posterior uncertainty. To address this challenge, one could simply compare the posteriors with and without a given feature included in the inference process. However, for large or nested feature sets, this would necessitate repeatedly estimating the posterior, which is computationally expensive or even prohibitive. Here, we provide a more efficient approach based on the SBI method neural likelihood estimation (NLE): We show that one can marginalize the trained surrogate likelihood post-hoc before inferring the posterior to assess the contribution of a feature. We demonstrate the usefulness of our method by identifying the most important features for inferring parameters of an example HH neuron model. Beyond neuroscience, our method is generally applicable to SBI workflows that rely on data features for inference used in other scientific fields.

## [Dynamics of SGD with Stochastic Polyak Stepsizes: Truly Adaptive Variants and Convergence to Exact Solution](#)

- Antonio Orvieto · Simon Lacoste-Julien · Nicolas Loizou
- abstract@[open-review](#): Recently Loizou et al. (2021), proposed and analyzed stochastic gradient descent (SGD) with stochastic Polyak stepsize (SPS). The proposed SPS comes with strong convergence guarantees and competitive performance; however, it has two main drawbacks when it is used in non-over-parameterized regimes: (i) It requires a priori knowledge of the optimal mini-batch losses, which are not available when the interpolation condition is not satisfied (e.g., regularized objectives), and (ii) it guarantees convergence only to a neighborhood of the solution. In this work, we study the dynamics and the convergence properties of SGD equipped with new variants of the stochastic Polyak stepsize and provide solutions to both drawbacks of the original SPS. We first show that a simple modification of the original SPS that uses lower bounds instead of the optimal function values can directly solve issue (i). On the other hand, solving issue (ii) turns out to be more challenging and leads us to valuable insights into the method's behavior. We show that if interpolation is not satisfied, the correlation between SPS and stochastic gradients introduces a bias, which effectively distorts the expectation of the gradient signal near minimizers, leading to non-convergence - even if the stepsize is scaled down during training. To fix this issue, we propose DecSPS, a novel modification of SPS, which guarantees convergence to the exact minimizer - without a priori knowledge of the problem parameters. For strongly-convex optimization problems, DecSPS is the first stochastic adaptive optimization method that converges to the exact solution without restrictive assumptions like bounded iterates/gradients.

## [A scalable tester for samplers](#)

- Yash Pote · Kuldeep S Meel
- abstract@[open-review](#): In this paper we study the problem of testing of constrained samplers over high-dimensional distributions with  $\$\\backslash varepsilon, \\eta, \\delta\$$  guarantees. Samplers are increasingly used in a wide range of safety-critical ML applications, and hence the testing problem has gained importance. For  $n$ -dimensional distributions, the existing state-of-the-art algorithm,  $\\mathsf{Barbarik2}$ , has a worst case query complexity of exponential in  $n$  and hence is not ideal for use in practice. Our primary contribution is an exponentially faster algorithm that has a query complexity linear in  $n$  and hence can easily scale to larger instances. We demonstrate our claim by implementing our algorithm and then comparing it against  $\\mathsf{Barbarik2}$ . Our experiments on the samplers  $\\mathsf{wUnigen3}$  and  $\\mathsf{wSTS}$ , find that  $\\mathsf{Pacoco}$  requires  $10 \\times$  fewer samples for  $\\mathsf{wUnigen3}$  and  $450 \\times$  fewer samples for  $\\mathsf{wSTS}$  as compared to  $\\mathsf{Barbarik2}$ .

## [Joint Learning of 2D-3D Weakly Supervised Semantic Segmentation](#)

- Hyeokjun Kweon · Kuk-Jin Yoon
- abstract@[open-review](#): The aim of weakly supervised semantic segmentation (WSSS) is to learn semantic segmentation without using dense annotations. WSSS has been intensively studied for 2D images and 3D point clouds. However, the existing WSSS studies have focused on a single domain, i.e. 2D or 3D, even when multi-domain data is available. In this paper, we propose a novel joint 2D-3D WSSS framework taking advantage of WSSS in different domains, using classification labels only. Via projection, we leverage the 2D class activation map as self-supervision to enhance the 3D semantic perception. Conversely, we exploit the similarity matrix of point cloud features for training the image classifier to achieve more precise 2D segmentation. In both directions, we devise a confidence-based scoring method to reduce the effect of inaccurate self-supervision. With extensive quantitative and qualitative experiments, we verify that the proposed joint WSSS framework effectively transfers the benefit of each domain to the other domain, and the resulting semantic segmentation performance is remarkably improved in both 2D and 3D domains. On ScanNetV2 benchmark, our framework significantly outperforms the prior WSSS approaches, suggesting a new research direction for WSSS.

## [Versatile Multi-stage Graph Neural Network for Circuit Representation](#)

- shuwen yang · Zhihao Yang · Dong Li · Yingxueff Zhang · Zhanhuang Zhang · Guojie Song · Jianye Hao
- abstract@[open-review](#): Due to the rapid growth in the scale of circuits and the desire for knowledge transfer from old designs to new ones, deep learning technologies have been widely exploited in Electronic Design Automation (EDA) to assist circuit design. In chip design cycles, we might encounter heterogeneous and diverse information sources, including the two most informative ones: the netlist and the design layout. However, handling each information source independently is sub-optimal. In this paper, we propose a novel way to integrate the multiple information sources under a unified heterogeneous graph named Circuit Graph, where topological and geometrical information is well integrated. Then, we propose Circuit GNN to fully utilize the features of vertices, edges as well as heterogeneous information during the message passing process. It is the first attempt to design a versatile circuit representation that is compatible across multiple EDA tasks and stages. Experiments on the two most representative prediction tasks in EDA show that our solution reaches state-of-the-art performance in both logic synthesis and global placement chip design stages. Besides, it achieves a 10x speed-up on congestion prediction compared to the state-of-the-art model.

## [Focal Modulation Networks](#)

- Jianwei Yang · Chunyuan Li · Xiyang Dai · Jianfeng Gao

- abstract@[open-review](#): In this work, we propose focal modulation network (FocalNet in short), where self-attention (SA) is completely replaced by a focal modulation module that is more effective and efficient for modeling token interactions. Focal modulation comprises three components: (i) hierarchical contextualization, implemented using a stack of depth-wise convolutional layers, to encode visual contexts from short to long ranges at different granularity levels, (ii) gated aggregation to selectively aggregate context features for each visual token (query) based on its content, and (iii) modulation or element-wise affine transformation to fuse the aggregated features into the query vector. Extensive experiments show that FocalNets outperform the state-of-the-art SA counterparts (e.g., Swin Transformers) with similar time and memory cost on the tasks of image classification, object detection, and semantic segmentation. Specifically, our FocalNets with tiny and base size achieve 82.3% and 83.9% top-1 accuracy on ImageNet-1K. After pretrained on ImageNet-22K, it attains 86.5% and 87.3% top-1 accuracy when finetuned with resolution 224 and 384, respectively. FocalNets exhibit remarkable superiority when transferred to downstream tasks. For object detection with Mask R-CNN, our FocalNet base trained with 1x already surpasses Swin trained with 3\$ \times \$ schedule (49.0 v.s. 48.5). For semantic segmentation with UperNet, FocalNet base evaluated at single-scale outperforms Swin evaluated at multi-scale (50.5 v.s. 49.7). These results render focal modulation a favorable alternative to SA for effective and efficient visual modeling in real-world applications.

## [Reinforcement Learning with Neural Radiance Fields](#)

- Danny Driess · Ingmar Schubert · Pete Florence · Yunzhu Li · Marc Toussaint
- abstract@[open-review](#): It is a long-standing problem to find effective representations for training reinforcement learning (RL) agents. This paper demonstrates that learning state representations with supervision from Neural Radiance Fields (NeRFs) can improve the performance of RL compared to other learned representations or even low-dimensional, hand-engineered state information. Specifically, we propose to train an encoder that maps multiple image observations to a latent space describing the objects in the scene. The decoder built from a latent-conditioned NeRF serves as the supervision signal to learn the latent space. An RL algorithm then operates on the learned latent space as its state representation. We call this NeRF-RL. Our experiments indicate that NeRF as supervision leads to a latent space better suited for the downstream RL tasks involving robotic object manipulations like hanging mugs on hooks, pushing objects, or opening doors.

## [On Computing Probabilistic Explanations for Decision Trees](#)

- Marcelo Arenas · Pablo Barceló · Miguel Romero Orth · Bernardo Subercaseaux
- abstract@[open-review](#): Formal XAI (explainable AI) is a growing area that focuses on computing explanations with mathematical guarantees for the decisions made by ML models. Inside formal XAI, one of the most studied cases is that of explaining the choices taken by decision trees, as they are traditionally deemed as one of the most interpretable classes of models. Recent work has focused on studying the computation of \emph{sufficient reasons}, a kind of explanation in which given a decision tree  $T$  and an instance  $e$ , one explains the decision  $T(e)$  by providing a subset  $\{e'\}$  of the features of  $e$  such that for any other instance  $e''$  compatible with  $e$  on  $\{e'\}$ , it holds that  $T(e'') = T(e)$ . One can argue, however, that sufficient reasons constitute a restrictive notion of explanation. For such a reason, the community has started to study their probabilistic counterpart, in which one requires that the probability of  $T(e'') = T(e)$  must be at least some value  $\delta \in (0, 1]$ , where  $e''$  is a random instance that is compatible with  $e$ . Our paper settles the computational complexity of  $\delta$ -sufficient-reasons over decision trees, showing that both (1) finding  $\delta$ -sufficient-reasons that are minimal in size, and (2) finding  $\delta$ -sufficient-reasons that are minimal inclusion-wise, do not admit polynomial-time algorithms (unless  $P = NP$ ). This is in stark contrast with the deterministic case ( $\delta = 1$ ) where inclusion-wise minimal sufficient-reasons are easy to compute. By doing this, we answer two open problems originally raised by Izza et al., and extend the hardness of explanations for Boolean circuits presented by Woldchen et al. to the more restricted case of decision trees. On the positive side, we identify structural restrictions of decision trees that make the problem tractable and show how SAT solvers might be able to tackle these problems in practical settings.

## [Amortised Inference in Structured Generative Models with Explaining Away](#)

- Changmin Yu · Hugo Soulat · Neil Burgess · Maneesh Sahani
- abstract@[open-review](#): A key goal of unsupervised learning is to go beyond density estimation and sample generation to reveal the structure inherent within observed data. Such structure can be expressed in the pattern of interactions between explanatory latent variables captured through a probabilistic graphical model. Although the learning of structured graphical models has a long history, much recent work in unsupervised modelling has instead emphasised flexible deep-network-based generation, either transforming independent latent generators to model complex data or assuming that distinct observed variables are derived from different latent nodes. Here, we extend the output of amortised variational inference to incorporate structured factors over multiple variables, able to capture the observation-induced posterior dependence between latents that results from "explaining away" and thus allow complex observations to depend on multiple nodes of a structured graph. We show that appropriately parameterised factors can be combined efficiently with variational message passing in elaborate graphical structures. We instantiate the framework based on Gaussian Process Factor Analysis models, and empirically evaluate its improvement over existing methods on synthetic data with known generative processes. We then fit the structured model to high-dimensional neural spiking time-series from the hippocampus of freely moving rodents, demonstrating that the model identifies latent signals that correlate with behavioural covariates.

## [AD-DROP: Attribution Driven Dropout for Robust Language Model Finetuning](#)

- Tao Yang · Jinghao Deng · Xiaojun Quan · Qifan Wang · Shaoliang Nie
- abstract@[open-review](#): Finetuning large pretrained language models on downstream tasks is apt to suffer from overfitting when limited training data is available. While dropout proves to be an effective antidote by randomly dropping a proportion of units, existing research has not examined its effect on the self-attention mechanism. In this paper, we investigate this problem through self-attention attribution and find that dropping attention positions with low attribution scores can accelerate training and increase the risk of overfitting. Motivated by this observation, we propose Attribution Driven Dropout (AD-DROP), which randomly discards high attribution positions to encourage the model to make predictions by relying more on low attribution positions to reduce overfitting. We also develop a cross-tuning strategy to alternate finetuning and AD-DROP to avoid dropping high attribution positions excessively. Extensive experiments on the GLUE benchmark show that AD-DROP not only effectively mitigates overfitting on small datasets but also leads to performance improvements on large datasets. These results confirm the success of AD-DROP as a strategic regularizer to prevent overfitting during finetuning.

## [Exploring Example Influence in Continual Learning](#)

- Qing Sun · Fan Lyu · Fanhua Shang · Wei Feng · Liang Wan
- abstract@[open-review](#): Continual Learning (CL) sequentially learns new tasks like human beings, with the goal to achieve better Stability (S, remembering past tasks) and Plasticity (P, adapting to new tasks). Due to the fact that past training data is not available, it is valuable to explore the influence difference on S and P among training examples, which may improve the learning pattern towards better SP. Inspired by Influence Function (IF), we first study example influence via adding perturbation to example weight and computing the influence derivation. To avoid the storage and calculation burden of Hessian inverse in neural networks, we propose a simple yet effective MetaSP algorithm to simulate the two key steps in the computation of IF and obtain the S- and P-aware example influence. Moreover, we propose to fuse two kinds of example influence by solving a Dual-Objective Optimization (DOO) problem, and obtain a fused influence towards SP Pareto optimality. The fused influence can be used to control the update of model and optimize the storage of rehearsal. Empirical results show that our algorithm significantly outperforms state-of-the-art methods on both task- and class-incremental benchmark CL datasets. We will make our code open-source.

## [Navigating Memory Construction by Global Pseudo-Task Simulation for Continual Learning](#)

- Yejia Liu · Wang Zhu · Shaolei Ren
- abstract@[open-review](#): Continual learning faces a crucial challenge of catastrophic forgetting. To address this challenge, experience replay (ER) that maintains a tiny subset of samples from previous tasks has been commonly used. Existing ER works usually focus on refining the learning objective for each task with a static memory construction policy. In this paper, we formulate the dynamic memory construction in ER as a combinatorial optimization problem, which aims at directly minimizing the global loss across all experienced tasks. We first apply three tactics to solve the problem in the offline setting as a starting point. To provide an approximate solution to this problem under the online continual learning setting, we further propose the Global Pseudo-task Simulation (GPS), which mimics future catastrophic forgetting of the current task by permutation. Our empirical results and analyses suggest that the GPS consistently improves accuracy across four commonly used vision benchmarks. We have also shown that our GPS can serve as the unified framework for integrating various memory construction policies in existing ER works.

## [SCINet: Time Series Modeling and Forecasting with Sample Convolution and Interaction](#)

- Minhao LIU · Ailing Zeng · Muxi Chen · Zhijian Xu · Qiuxia LAI · Lingna Ma · Qiang Xu
- abstract@[open-review](#): One unique property of time series is that the temporal relations are largely preserved after downsampling into two sub-sequences. By taking advantage of this property, we propose a novel neural network architecture that conducts sample convolution and interaction for temporal modeling and forecasting, named SCINet. Specifically, SCINet is a recursive downsample-convolve-interact architecture. In each layer, we use multiple convolutional filters to extract distinct yet valuable temporal features from the downsampled sub-sequences or features. By combining these rich features aggregated from multiple resolutions, SCINet effectively models time series with complex temporal dynamics. Experimental results show that SCINet achieves significant forecasting accuracy improvements over both existing convolutional models and Transformer-based solutions across various real-world time series forecasting datasets. Our codes and data are available at <https://anonymous.4open.science/r/SCINet-2588>.

## [Concentration of Data Encoding in Parameterized Quantum Circuits](#)

- Guangxi Li · Ruilin Ye · Xuanqiang Zhao · Xin Wang
- abstract@[open-review](#): Variational quantum algorithms have been acknowledged as the leading strategy to realize near-term quantum advantages in meaningful tasks, including machine learning and optimization. When applied to tasks involving classical data, such algorithms generally begin with data encoding circuits and train quantum neural networks (QNNs) to minimize target functions. Although QNNs have been widely studied to improve these algorithms' performance on practical tasks, there is a gap in systematically understanding the influence of data encoding on the eventual performance. In this paper, we make progress in filling this gap by considering the common data encoding strategies based on parameterized quantum circuits. We prove that, under reasonable assumptions, the distance between the average encoded state and the maximally mixed state could be explicitly upper-bounded with respect to the width and depth of the encoding circuit. This result in particular implies that the average encoded state will concentrate on the maximally mixed state at an exponential speed on depth. Such concentration seriously limits the capabilities of quantum classifiers, and strictly restricts the distinguishability of encoded states from a quantum information perspective. To support our findings, we numerically verify these results on both synthetic and public data sets. Our results highlight the significance of quantum data encoding and may shed light on future encoding strategies.

## [Imbalance Trouble: Revisiting Neural-Collapse Geometry](#)

- Christos Thrampoulidis · Ganesh Ramachandra Kini · Vala Vakilian · Tina Behnia
- abstract@[open-review](#): Neural Collapse refers to the remarkable structural properties characterizing the geometry of class embeddings and classifier weights, found by deep nets when trained beyond zero training error. However, this characterization only holds for balanced data. Here we thus ask whether it can be made invariant to class imbalances. Towards this end, we adopt the unconstrained feature model (UFM), a recent theoretical model for studying neural collapse, and introduce \$text{\texttt{Simplex-Encoded-Labels Interpolation}}\$ (SELI) as an invariant characterization of the neural collapse phenomenon. Specifically, we prove for the UFM with cross-entropy loss and vanishing regularization that, irrespective of class imbalances, the embeddings and classifiers always interpolate a simplex-encoded label matrix and that their individual geometries are determined by the SVD factors of this same label matrix. We then present extensive experiments on synthetic and real datasets that confirm convergence to the SELI geometry. However, we caution that convergence worsens with increasing imbalances. We theoretically support this finding by showing that unlike the balanced case, when minorities are present, ridge-regularization plays a critical role in tweaking the geometry. This defines new questions and motivates further investigations into the impact of class imbalances on the rates at which first-order methods converge to their asymptotically preferred solutions.

## [RCNNs Learn Succinct Learning Algorithms in Polynomial Time](#)

- Surbhi Goel · Cyril Zhang · Sham Kakade · Adam Kalai
- abstract@[open-review](#): Neural Networks (NNs) struggle to efficiently learn certain problems, such as parity problems, even when there are simple learning algorithms for those problems. Can NNs discover learning algorithms on their own? We exhibit a NN architecture that, in polynomial time, learns as well as any efficient learning algorithm describable by a constant-sized learning algorithm. For example, on parity problems, the NN learns as well as row reduction, an efficient algorithm that can be succinctly described. Our architecture combines both recurrent weight-sharing between layers and convolutional weight-sharing to reduce the number of \textit{parameters} down to a constant, even though the network itself may have trillions of nodes. While in practice the constants in our analysis are too large to be directly meaningful, our work suggests that the synergy of Recurrent and Convolutional NNs (RCNNs) may be more powerful than either alone.

## [Hidden Progress in Deep Learning: SGD Learns Parities Near the Computational Limit](#)

- Boaz Barak · Benjamin Edelman · Surbhi Goel · Sham Kakade · Eran Malach · Cyril Zhang
- abstract@[open-review](#): There is mounting empirical evidence of emergent phenomena in the capabilities of deep learning methods as we scale up datasets, model sizes, and training times. While there are some accounts of how these resources modulate statistical capacity, far less is known about their effect on the computational problem of model training. This work conducts such an exploration through the lens of learning  $k\$$ -sparse parities of  $n\$$  bits, a canonical family of problems which pose theoretical computational barriers. In this setting, we find that neural networks exhibit surprising phase transitions when scaling up dataset size and running time. In particular, we demonstrate empirically that with standard training, a variety of architectures learn sparse parities with  $n^{O(k)}$  examples, with loss (and error) curves abruptly dropping after  $n^{O(k)}$  iterations. These positive results nearly match known SQ lower bounds, even without an explicit sparsity-promoting prior. We elucidate the mechanisms of these phenomena with a theoretical analysis: we find that the phase transition in performance is not due to SGD "stumbling in the dark" until it finds the hidden set of features; instead, we show that SGD gradually amplifies a Fourier gap in the population gradient.

## [Low-Rank Modular Reinforcement Learning via Muscle Synergy](#)

- Heng Dong · Tonghan Wang · Chongjie Zhang
- abstract@[open-review](#): Modular Reinforcement Learning (RL) decentralizes the control of multi-joint robots by learning policies for each actuator. Previous work on modular RL has proven its ability to control morphologically different agents with a shared actuator policy. However, with the increase in the Degree of Freedom (DoF) of robots, training a morphology-generalizable modular controller becomes exponentially difficult. Motivated by the way

the human central nervous system controls numerous muscles, we propose a Synergy-Oriented LeARning (SOLAR) framework that exploits the redundant nature of DoF in robot control. Actuators are grouped into synergies by an unsupervised learning method, and a synergy action is learned to control multiple actuators in synchrony. In this way, we achieve a low-rank control at the synergy level. We extensively evaluate our method on a variety of robot morphologies, and the results show its superior efficiency and generalizability, especially on robots with a large DoF like Humanoids++ and UNIMALs.

## [DreamShard: Generalizable Embedding Table Placement for Recommender Systems](#)

- Daochen Zha · Louis Feng · Qiaoyu Tan · Zirui Liu · Kwei-Herng Lai · Bhargav Bhushanam · Yuandong Tian · Arun Kejariwal · Xia Hu
- abstract@[open-review](#): We study embedding table placement for distributed recommender systems, which aims to partition and place the tables on multiple hardware devices (e.g., GPUs) to balance the computation and communication costs. Although prior work has explored learning-based approaches for the device placement of computational graphs, embedding table placement remains to be a challenging problem because of 1) the operation fusion of embedding tables, and 2) the generalizability requirement on unseen placement tasks with different numbers of tables and/or devices. To this end, we present DreamShard, a reinforcement learning (RL) approach for embedding table placement. DreamShard achieves the reasoning of operation fusion and generalizability with 1) a cost network to directly predict the costs of the fused operation, and 2) a policy network that is efficiently trained on an estimated Markov decision process (MDP) without real GPU execution, where the states and the rewards are estimated with the cost network. Equipped with sum and max representation reductions, the two networks can directly generalize to any unseen tasks with different numbers of tables and/or devices without fine-tuning. Extensive experiments show that DreamShard substantially outperforms the existing human expert and RNN-based strategies with up to 19% speedup over the strongest baseline on large-scale synthetic tables and our production tables. The code will be open-sourced.

## [Expediting Large-Scale Vision Transformer for Dense Prediction without Fine-tuning](#)

- WEICONG LIANG · YUHUI YUAN · Henghui Ding · Xiao Luo · Weihong Lin · Ding Jia · Zheng Zhang · Chao Zhang · Han Hu
- abstract@[open-review](#): Vision transformers have recently achieved competitive results across various vision tasks but still suffer from heavy computation costs when processing a large number of tokens. Many advanced approaches have been developed to reduce the total number of tokens in the large-scale vision transformers, especially for image classification tasks. Typically, they select a small group of essential tokens according to their relevance with the `[text{class}]` token, then fine-tune the weights of the vision transformer. Such fine-tuning is less practical for dense prediction due to the much heavier computation and GPU memory cost than image classification. In this paper, we focus on a more challenging problem, *i.e.*, accelerating large-scale vision transformers for dense prediction without any additional re-training or fine-tuning. In response to the fact that high-resolution representations are necessary for dense prediction, we present two non-parametric operators, a `\emph{token clustering layer}` to decrease the number of tokens and a `\emph{token reconstruction layer}` to increase the number of tokens. The following steps are performed to achieve this: (i) we use the token clustering layer to cluster the neighboring tokens together, resulting in low-resolution representations that maintain the spatial structures; (ii) we apply the following transformer layers only to these low-resolution representations or clustered tokens; and (iii) we use the token reconstruction layer to re-create the high-resolution representations from the refined low-resolution representations. The results obtained by our method are promising on five dense prediction tasks including object detection, semantic segmentation, panoptic segmentation, instance segmentation, and depth estimation. Accordingly, our method accelerates \$40\%\uparrow\$ FPS and saves \$30\%\downarrow\$ GFLOPs of ``Segmener+ViT-L/\$16\\$'' while maintaining \$99.5\%\$ of the performance on ADE\$20\\$K without fine-tuning the official weights.

## [Stimulative Training of Residual Networks: A Social Psychology Perspective of Loafing](#)

- Peng Ye · Shengji Tang · Baopu Li · Tao Chen · Wanli Ouyang
- abstract@[open-review](#): Residual networks have shown great success and become indispensable in today's deep models. In this work, we aim to re-investigate the training process of residual networks from a novel social psychology perspective of loafing, and further propose a new training strategy to strengthen the performance of residual networks. As residual networks can be viewed as ensembles of relatively shallow networks (*i.e.*, `\textit{unraveled view}`) in prior works, we also start from such view and consider that the final performance of a residual network is co-determined by a group of sub-networks. Inspired by the social loafing problem of social psychology, we find that residual networks invariably suffer from similar problem, where sub-networks in a residual network are prone to exert less effort when working as part of the group compared to working alone. We define this previously overlooked problem as `\textit{network loafing}`. As social loafing will ultimately cause the low individual productivity and the reduced overall performance, network loafing will also hinder the performance of a given residual network and its sub-networks. Referring to the solutions of social psychology, we propose `\textit{stimulative training}`, which randomly samples a residual sub-network and calculates the KL-divergence loss between the sampled sub-network and the given residual network, to act as extra supervision for sub-networks and make the overall goal consistent. Comprehensive empirical results and theoretical analyses verify that stimulative training can well handle the loafing problem, and improve the performance of a residual network by improving the performance of its sub-networks.

## [DENSE: Data-Free One-Shot Federated Learning](#)

- Jie Zhang · Chen Chen · Bo Li · Lingjuan Lyu · Shuang Wu · Shouhong Ding · Chunhua Shen · Chao Wu
- abstract@[open-review](#): One-shot Federated Learning (FL) has recently emerged as a promising approach, which allows the central server to learn a model in a single communication round. Despite the low communication cost, existing one-shot FL methods are mostly impractical or face inherent limitations, *e.g.* a public dataset is required, %poor performance of the global model, clients' models are homogeneous, and additional data/model information need to be uploaded. To overcome these issues, we propose a novel two-stage `\textbf{D}ata-fre\textbf{E}x\textbf{N}e\textbf{S}hot federated \textbf{E}arning` (DENSE) framework, which trains the global model by a data generation stage and a model distillation stage. DENSE is a practical one-shot FL method that can be applied in reality due to the following advantages:(1) DENSE requires no additional information compared with other methods (except the model parameters) to be transferred between clients and the server;(2) DENSE does not require any auxiliary dataset for training;(3) DENSE considers model heterogeneity in FL, *i.e.* different clients can have different model architectures.Experiments on a variety of real-world datasets demonstrate the superiority of our method.For example, DENSE outperforms the best baseline method Fed-ADI by 5.08% on CIFAR10 dataset. Our code will soon be available.

## [Accelerated Projected Gradient Algorithms for Sparsity Constrained Optimization Problems](#)

- Jan Harold Alcantara · Ching-pei Lee
- abstract@[open-review](#): We consider the projected gradient algorithm for the nonconvex best subset selection problem that minimizes a given empirical loss function under an  $\ell_0$ -norm constraint. Through decomposing the feasible set of the given sparsity constraint as a finite union of linear subspaces, we present two acceleration schemes with global convergence guarantees, one by same-space extrapolation and the other by subspace identification. The former fully utilizes the problem structure to greatly accelerate the optimization speed with only negligible additional cost. The latter leads to a two-stage meta-algorithm that first uses classical projected gradient iterations to identify the correct subspace containing an optimal solution, and then switches to a highly-efficient smooth optimization method in the identified subspace to attain superlinear convergence. Experiments demonstrate that the proposed accelerated algorithms are magnitudes faster than their non-accelerated counterparts as well as the state of the art.

## [Meta-Learning with Self-Improving Momentum Target](#)

- Jihoon Tack · Jongjin Park · Hankook Lee · Jaeho Lee · Jinwoo Shin

- abstract@[open-review](#): The idea of using a separately trained target model (or teacher) to improve the performance of the student model has been increasingly popular in various machine learning domains, and meta-learning is no exception; a recent discovery shows that utilizing task-wise target models can significantly boost the generalization performance. However, obtaining a target model for each task can be highly expensive, especially when the number of tasks for meta-learning is large. To tackle this issue, we propose a simple yet effective method, coined Self-improving Momentum Target (SiMT). SiMT generates the target model by adapting from the temporal ensemble of the meta-learner, i.e., the momentum network. This momentum network and its task-specific adaptations enjoy a favorable generalization performance, enabling self-improving of the meta-learner through knowledge distillation. Moreover, we found that perturbing parameters of the meta-learner, e.g., dropout, further stabilize this self-improving process by preventing fast convergence of the distillation loss during meta-training. Our experimental results demonstrate that SiMT brings a significant performance gain when combined with a wide range of meta-learning methods under various applications, including few-shot regression, few-shot classification, and meta-reinforcement learning.

## [Quasi-Newton Methods for Saddle Point Problems](#)

- Chengchang Liu Â· Luo Luo
- abstract@[open-review](#): This paper studies quasi-Newton methods for strongly-convex-strongly-concave saddle point problems. We propose random Broyden family updates, which have explicit local superlinear convergence rate of  $\mathcal{O}(\big(1-1/(n\kappa^2)\big)^{k(k-1)/2})$ , where  $n$  is the dimension of the problem,  $\kappa$  is the condition number and  $k$  is the number of iterations. The design and analysis of proposed algorithm are based on estimating the square of indefinite Hessian matrix, which is different from classical quasi-Newton methods in convex optimization. We also present two specific Broyden family algorithms with BFGS-type and SR1-type updates, which enjoy the faster local convergence rate of  $\mathcal{O}(\big(1-1/n\big)^{k(k-1)/2})$ . Our numerical experiments show proposed algorithms outperform classical first-order methods.

## [Training language models to follow instructions with human feedback](#)

- Long Ouyang Â· Jeffrey Wu Â· Xu Jiang Â· Diogo Almeida Â· Carroll Wainwright Â· Pamela Mishkin Â· Chong Zhang Â· Sandhini Agarwal Â· Katarina Slama Â· Alex Ray Â· John Schulman Â· Jacob Hilton Â· Fraser Kelton Â· Luke Miller Â· Maddie Simens Â· Amanda Askell Â· Peter Welinder Â· Paul Christiano Â· Jan Leike Â· Ryan Lowe
- abstract@[open-review](#): Making language models bigger does not inherently make them better at following a user's intent. For example, large language models can generate outputs that are untruthful, toxic, or simply not helpful to the user. In other words, these models are not aligned with their users. In this paper, we show an avenue for aligning language models with user intent on a wide range of tasks by fine-tuning with human feedback. Starting with a set of labeler-written prompts and prompts submitted through a language model API, we collect a dataset of labeler demonstrations of the desired model behavior, which we use to fine-tune GPT-3 using supervised learning. We then collect a dataset of rankings of model outputs, which we use to further fine-tune this supervised model using reinforcement learning from human feedback. We call the resulting models InstructGPT. In human evaluations on our prompt distribution, outputs from the 1.3B parameter InstructGPT model are preferred to outputs from the 175B GPT-3, despite having 100x fewer parameters. Moreover, InstructGPT models show improvements in truthfulness and reductions in toxic output generation while having minimal performance regressions on public NLP datasets. Even though InstructGPT still makes simple mistakes, our results show that fine-tuning with human feedback is a promising direction for aligning language models with human intent.

## [Deterministic Langevin Monte Carlo with Normalizing Flows for Bayesian Inference](#)

- Uros Seljak Â· Richard Grumitt Â· Biwei Dai
- abstract@[open-review](#): We propose a general purpose Bayesian inference algorithm for expensive likelihoods, replacing the stochastic term in Langevin equation with a deterministic density gradient term. Particle density is evaluated from the current particle positions using a Normalizing Flow (NF), which is differentiable and has good generalization properties in high dimensions. We additionally take advantage of NF preconditioning and NF based Metropolis-Hastings updates for a faster convergence. We show on various examples that the method is competitive against the state of the art sampling methods.

## [Batch size-invariance for policy optimization](#)

- Jacob Hilton Â· Karl Cobbe Â· John Schulman
- abstract@[open-review](#): We say an algorithm is batch size-invariant if changes to the batch size can largely be compensated for by changes to other hyperparameters. Stochastic gradient descent is well-known to have this property at small batch sizes, via the learning rate. However, some policy optimization algorithms (such as PPO) do not have this property, because of how they control the size of policy updates. In this work we show how to make these algorithms batch size-invariant. Our key insight is to decouple the proximal policy (used for controlling policy updates) from the behavior policy (used for off-policy corrections). Our experiments help explain why these algorithms work, and additionally show how they can make more efficient use of stale data.

## [Rate-Distortion Theoretic Bounds on Generalization Error for Distributed Learning](#)

- Milad Sefidgaran Â· Romain Chor Â· Abdellatif Zaidi
- abstract@[open-review](#): In this paper, we use tools from rate-distortion theory to establish new upper bounds on the generalization error of statistical distributed learning algorithms. Specifically, there are  $K$  clients whose individually chosen models are aggregated by a central server. The bounds depend on the compressibility of each client's algorithm while keeping other clients' algorithms un-compressed, and leveraging the fact that small changes in each local model change the aggregated model by a factor of only  $1/K$ . Adopting a recently proposed approach by Sefidgaran et al., and extending it suitably to the distributed setting, enables smaller rate-distortion terms which are shown to translate into tighter generalization bounds. The bounds are then applied to the distributed support vector machines (SVM), suggesting that the generalization error of the distributed setting decays faster than that of the centralized one with a factor of  $\mathcal{O}(\sqrt{\log(K)/K})$ . This finding is validated also experimentally. A similar conclusion is obtained for a multiple-round federated learning setup where each client uses stochastic gradient Langevin dynamics (SGLD).

## [Multi-Objective Bayesian Optimization with Pareto Set Learning](#)

- Xi Lin Â· Zhiyuan Yang Â· Xiaoyuan Zhang Â· Qingfu Zhang
- abstract@[open-review](#): Expensive multi-objective optimization problems can be found in many real-world applications, where their objective function evaluations involve expensive computation and/or physical experiments. It is desirable to obtain an approximate Pareto front with a small evaluation budget. Multi-objective Bayesian optimization (MOBO) has been widely used for finding a finite set of Pareto optimal solutions to this problem. However, it is well-known that the whole Pareto set could have infinite optimal solutions. The structural properties of the Pareto set are not well utilized in existing MOBO methods, and the finite-set approximation may not contain the most preferred solution(s) for decision-makers. This paper develops a simple yet efficient Pareto set learning method to approximate the whole Pareto set for MOBO. We design a simple yet powerful acquisition search method with the learned Pareto set, which naturally supports batch evaluation. In addition, with our proposed model, decision-makers can readily explore any trade-off area in the approximate Pareto set for flexible decision-making. Experimental results on different synthetic and real-world problems demonstrate the effectiveness and efficiency of our proposed method.

## Understanding Square Loss in Training Overparametrized Neural Network Classifiers

- Tianyang Hu · Jun WANG · Wenjia Wang · Zhenguo Li
- abstract@[open-review](#): Deep learning has achieved many breakthroughs in modern classification tasks. Numerous architectures have been proposed for different data structures but when it comes to the loss function, the cross-entropy loss is the predominant choice. Recently, several alternative losses have seen revived interests for deep classifiers. In particular, empirical evidence seems to promote square loss but a theoretical justification is still lacking. In this work, we contribute to the theoretical understanding of square loss in classification by systematically investigating how it performs for overparametrized neural networks in the neural tangent kernel (NTK) regime. Interesting properties regarding the generalization error, robustness, and calibration error are revealed. We consider two cases, according to whether classes are separable or not. In the general non-separable case, fast convergence rate is established for both misclassification rate and calibration error. When classes are separable, the misclassification rate improves to be exponentially fast. Further, the resulting margin is proven to be lower bounded away from zero, providing theoretical guarantees for robustness. We expect our findings to hold beyond the NTK regime and translate to practical settings. To this end, we conduct extensive empirical studies on practical neural networks, demonstrating the effectiveness of square loss in both synthetic low-dimensional data and real image data. Comparing to cross-entropy, square loss has comparable generalization error but noticeable advantages in robustness and model calibration.

## A Win-win Deal: Towards Sparse and Robust Pre-trained Language Models

- Yuanxin Liu · Fandong Meng · Zheng Lin · Jiangnan Li · Peng Fu · Yanan Cao · Weiping Wang · Jie Zhou
- abstract@[open-review](#): Despite the remarkable success of pre-trained language models (PLMs), they still face two challenges: First, large-scale PLMs are inefficient in terms of memory footprint and computation. Second, on the downstream tasks, PLMs tend to rely on the dataset bias and struggle to generalize to out-of-distribution (OOD) data. In response to the efficiency problem, recent studies show that dense PLMs can be replaced with sparse subnetworks without hurting the performance. Such subnetworks can be found in three scenarios: 1) the fine-tuned PLMs, 2) the raw PLMs and then fine-tuned in isolation, and even inside 3) PLMs without any parameter fine-tuning. However, these results are only obtained in the in-distribution (ID) setting. In this paper, we extend the study on PLMs subnetworks to the OOD setting, investigating whether sparsity and robustness to dataset bias can be achieved simultaneously. To this end, we conduct extensive experiments with the pre-trained BERT model on three natural language understanding (NLU) tasks. Our results demonstrate that \textbf{sparse} and robust subnetworks (SRNets) can consistently be found in BERT\}, across the aforementioned three scenarios, using different training and compression methods. Furthermore, we explore the upper bound of SRNets using the OOD information and show that \textbf{there exist sparse and almost unbiased BERT subnetworks}. Finally, we refine the SRNets searching process in terms of efficiency and performance, which involves: 1) the appropriate timing to start searching SRNets during full BERT fine-tuning, and 2) how to identify SRNets at high sparsity. Our codes will be released on publication.

## DASCO: Dual-Generator Adversarial Support Constrained Offline Reinforcement Learning

- Quan Vuong · Aviral Kumar · Sergey Levine · Yevgen Chebotar
- abstract@[open-review](#): In offline RL, constraining the learned policy to remain close to the data is essential to prevent the policy from outputting out-of-distribution (OOD) actions with erroneously overestimated values. In principle, generative adversarial networks (GAN) can provide an elegant solution to do so, with the discriminator directly providing a probability that quantifies distributional shift. However, in practice, GAN-based offline RL methods have not outperformed alternative approaches, perhaps because the generator is trained to both fool the discriminator and maximize return - two objectives that are often at odds with each other. In this paper, we show that the issue of conflicting objectives can be resolved by training two generators: one that maximizes return, with the other capturing the "remainder" of the data distribution in the offline dataset, such that the mixture of the two is close to the behavior policy. We show that not only does having two generators enable an effective GAN-based offline RL method, but also approximates a support constraint, where the policy does not need to match the entire data distribution, but only the slice of the data that leads to high long term performance. We name our method DASCO, for Dual-Generator Adversarial Support Constrained Offline RL. On benchmark tasks that require learning from sub-optimal data, DASCO significantly outperforms prior methods that enforce distribution constraint.

## A Differentially Private Linear-Time fPTAS for the Minimum Enclosing Ball Problem

- Bar Mahpud · Or Sheffet
- abstract@[open-review](#): The Minimum Enclosing Ball (MEB) problem is one of the most fundamental problems in clustering, with applications in operations research, statistic and computational geometry. In this works, we give the first differentially private (DP) fPTAS for the Minimum Enclosing Ball problem, improving both on the runtime and the utility bound of the best known DP-PTAS for the problem, of Ghazi et al (2020). Given \$n\$ points in \$\mathbb{R}^d\$ that are covered by the ball \$B(\theta\_{opt}, r\_{opt})\$, our simple iterative DP-algorithm returns a ball \$B(\theta, r)\$ where \$r \leq (1+\gamma)r\_{opt}\$ and which leaves at most \$\tilde{O}(\frac{\sqrt{d}}{\gamma^2\epsilon})\$ points uncovered in \$\tilde{O}(n/\gamma^2)\$-time. We also give a local-model version of our algorithm, that leaves at most \$\tilde{O}(\frac{\sqrt{nd}}{\gamma^2\epsilon})\$ points uncovered, improving on the \$n^{0.67}\$-bound of Nissim and Stemmer (2018) (at the expense of other parameters). In addition, we test our algorithm empirically and discuss future open problems.

## Joint Entropy Search for Multi-Objective Bayesian Optimization

- Ben Tu · Axel Gandy · Nikolas Kantas · Behrang Shafei
- abstract@[open-review](#): Many real-world problems can be phrased as a multi-objective optimization problem, where the goal is to identify the best set of compromises between the competing objectives. Multi-objective Bayesian optimization (BO) is a sample efficient strategy that can be deployed to solve these vector-valued optimization problems where access is limited to a number of noisy objective function evaluations. In this paper, we propose a novel information-theoretic acquisition function for BO called Joint Entropy Search (JES), which considers the joint information gain for the optimal set of inputs and outputs. We present several analytical approximations to the JES acquisition function and also introduce an extension to the batch setting. We showcase the effectiveness of this new approach on a range of synthetic and real-world problems in terms of the hypervolume and its weighted variants.

## Theoretically Provable Spiking Neural Networks

- Shao-Qun Zhang · Zhi-Hua Zhou
- abstract@[open-review](#): Spiking neural networks have attracted increasing attention in recent years due to their potential of handling time-dependent data. Many algorithms and techniques have been developed; however, theoretical understandings of many aspects of spiking neural networks are far from clear. A recent work [Zhang and Zhou, 2021] disclosed that typical spiking neural networks could hardly work on spatio-temporal data due to their bifurcation dynamics and suggested that \textit{self-connection} has to be added. In this paper, we theoretically investigate the approximation powers and computational efficiency of spiking neural networks with self connections, and show that the self-connection structure enables spiking neural networks to approximate continuous dynamical systems within polynomial parameters and time complexities. Our theoretical results may shed some insights on developing provable and sound spiking neural networks.

## Follow-the-Perturbed-Leader for Adversarial Markov Decision Processes with Bandit Feedback

- Yan Dai · Haipeng Luo · Liyu Chen

- abstract@[open-review](#): We consider regret minimization for Adversarial Markov Decision Processes (AMDPs), where the loss functions are changing over time and adversarially chosen, and the learner only observes the losses for the visited state-action pairs (i.e., bandit feedback). While there has been a surge of studies on this problem using Online-Mirror-Descent (OMD) methods, very little is known about the Follow-the-Perturbed-Leader (FTPL) methods, which are usually computationally more efficient and also easier to implement since it only requires solving an offline planning problem. Motivated by this, we take a closer look at FTPL for learning AMDPs, starting from the standard episodic finite-horizon setting. We find some unique and intriguing difficulties in the analysis and propose a workaround to eventually show that FTPL is also able to achieve near-optimal regret bounds in this case. More importantly, we then find two significant applications: First, the analysis of FTPL turns out to be readily generalizable to delayed bandit feedback with order-optimal regret, while OMD methods exhibit extra difficulties (Jin et al., 2022). Second, using FTPL, we also develop the first no-regret algorithm for learning communicating AMDPs in the infinite-horizon setting with bandit feedback and stochastic transitions. Our algorithm is efficient assuming access to an offline planning oracle, while even for the easier full-information setting, the only existing algorithm (Chandrasekaran and Tewari, 2021) is computationally inefficient.

## [Generative Status Estimation and Information Decoupling for Image Rain Removal](#)

- Di Lin Â· Xin WANG Â· Jia Shen Â· Renjie Zhang Â· Ruonan Liu Â· Miaoqiu Wang Â· Wuyuan Xie Â· Qing Guo Â· Ping Li
- abstract@[open-review](#): Image rain removal requires the accurate separation between the pixels of the rain streaks and object textures. But the confusing appearances of rains and objects lead to the misunderstanding of pixels, thus remaining the rain streaks or missing the object details in the result. In this paper, we propose SEIDNet equipped with the generative Status Estimation and Information Decoupling for rain removal. In the status estimation, we embed the pixel-wise statuses into the status space, where each status indicates a pixel of the rain or object. The status space allows sampling multiple statuses for a pixel, thus capturing the confusing rain or object. In the information decoupling, we respect the pixel-wise statuses, decoupling the appearance information of rain and object from the pixel. Based on the decoupled information, we construct the kernel space, where multiple kernels are sampled for the pixel to remove the rain and recover the object appearance. We evaluate SEIDNet on the public datasets, achieving state-of-the-art performances of image rain removal. The experimental results also demonstrate the generalization of SEIDNet, which can be easily extended to achieve state-of-the-art performances on other image restoration tasks (e.g., snow, haze, and shadow removal).

## [Off-Policy Evaluation for Episodic Partially Observable Markov Decision Processes under Non-Parametric Models](#)

- Rui Miao Â· Zhengling Qi Â· Xiaoke Zhang
- abstract@[open-review](#): We study the problem of off-policy evaluation (OPE) for episodic Partially Observable Markov Decision Processes (POMDPs) with continuous states. Motivated by the recently proposed proximal causal inference framework, we develop a non-parametric identification result for estimating the policy value via a sequence of so-called V-bridge functions with the help of time-dependent proxy variables. We then develop a fitted-Q-evaluation-type algorithm to estimate V-bridge functions recursively, where a non-parametric instrumental variable (NPIV) problem is solved at each step. By analyzing this challenging sequential NPIV estimation, we establish the finite-sample error bounds for estimating the V-bridge functions and accordingly that for evaluating the policy value, in terms of the sample size, length of horizon and so-called (local) measure of ill-posedness at each step. To the best of our knowledge, this is the first finite-sample error bound for OPE in POMDPs under non-parametric models.

## [Exploring the Limits of Domain-Adaptive Training for Detoxifying Large-Scale Language Models](#)

- Boxin Wang Â· Wei Ping Â· Chaowei Xiao Â· Peng Xu Â· Mostofa Patwary Â· Mohammad Shoeybi Â· Bo Li Â· Anima Anandkumar Â· Bryan Catanzaro
- abstract@[open-review](#): Pre-trained language models (LMs) are shown to easily generate toxic language. In this work, we explore domain-adaptive training to reduce the toxicity of language models. We conduct this study on three dimensions: training corpus, model size, and parameter efficiency. For the training corpus, we first propose to leverage the generative power of LMs and generate nontoxic datasets for domain-adaptive training, which is shown to be more data-efficient than using a curated pre-training corpus. We demonstrate that the self-generation method consistently outperforms the existing baselines across various model sizes on both automatic and human evaluations, even when it uses a 1/3 smaller training corpus. We then comprehensively study detoxifying LMs with parameter sizes ranging from 126M up to 530B~(3x larger than GPT-3), a scale that has never been studied before. We find that i) large LMs have similar toxicity levels as smaller ones given the same pre-training corpus, and ii) large LMs require more endeavor to detoxify. We also explore parameter-efficient training methods for detoxification. We demonstrate that adding and training adapter-only layers in LMs not only saves a lot of parameters but also achieves a better trade-off between toxicity and perplexity than whole model adaptation for the large-scale models.

## [BiT: Robustly Binarized Multi-distilled Transformer](#)

- Zechun Liu Â· Barlas Oguz Â· Aasish Pappu Â· Lin Xiao Â· Scott Yih Â· Meng Li Â· Raghuraman Krishnamoorthi Â· Yashar Mehdad
- abstract@[open-review](#): Modern pre-trained transformers have rapidly advanced the state-of-the-art in machine learning, but have also grown in parameters and computational complexity, making them increasingly difficult to deploy in resource-constrained environments. Binarization of the weights and activations of the network can significantly alleviate these issues, however, is technically challenging from an optimization perspective. In this work, we identify a series of improvements that enables binary transformers at a much higher accuracy than what was possible previously. These include a two-set binarization scheme, a novel elastic binary activation function with learned parameters, and a method to quantize a network to its limit by successively distilling higher precision models into lower precision students. These approaches allow for the first time, fully binarized transformer models that are at a practical level of accuracy, approaching a full-precision BERT baseline on the GLUE language understanding benchmark within as little as 5.9%. Code and models are available at: <https://github.com/facebookresearch/bit>.

## [EfficientViT: Vision Transformers at MobileNet Speed](#)

- Yanyu Li Â· Geng Yuan Â· Yang Wen Â· Ju Hu Â· Georgios Evangelidis Â· Sergey Tulyakov Â· Yanzhi Wang Â· Jian Ren
- abstract@[open-review](#): Vision Transformers (ViT) have shown rapid progress in computer vision tasks, achieving promising results on various benchmarks. However, due to the massive number of parameters and model design, e.g., attention mechanism, ViT-based models are generally times slower than lightweight convolutional networks. Therefore, the deployment of ViT for real-time applications is particularly challenging, especially on resource-constrained hardware such as mobile devices. Recent efforts try to reduce the computation complexity of ViT through network architecture search or hybrid design with MobileNet block, yet the inference speed is still unsatisfactory. This leads to an important question: can transformers run as fast as MobileNet while obtaining high performance? To answer this, we first revisit the network architecture and operators used in ViT-based models and identify inefficient designs. Then we introduce a dimension-consistent pure transformer (without MobileNet blocks) as design paradigm. Finally, we perform latency-driven slimming to get a series of final models dubbed EfficientViT. Extensive experiments show the superiority of EfficientViT in performance and speed on mobile devices. Our fastest model, EfficientViT-L1, achieves 79.2% top-1 accuracy on ImageNet-1K with only 1.6 ms inference latency on iPhone 12 (compiled with CoreML), which is even a bit faster than MobileNetV2 (1.7 ms, 71.8% top-1), and our largest model, EfficientViT-L7, obtains 83.3% accuracy with only 7.0 ms latency. Our work proves that properly designed transformers can reach extremely low latency on mobile devices while maintaining high performance.

## [Improving 3D-aware Image Synthesis with A Geometry-aware Discriminator](#)

- Zifan Shi Â· Yinghao Xu Â· Yujun Shen Â· Deli Zhao Â· Qifeng Chen Â· Dit-Yan Yeung

- abstract@[open-review](#): 3D-aware image synthesis aims at learning a generative model that can render photo-realistic 2D images while capturing decent underlying 3D shapes. A popular solution is to adopt the generative adversarial network (GAN) and replace the generator with a 3D renderer, where volume rendering with neural radiance field (NeRF) is commonly used. Despite the advancement of synthesis quality, existing methods fail to obtain moderate 3D shapes. We argue that, considering the two-player game in the formulation of GANs, only making the generator 3D-aware is not enough. In other words, displacing the generative mechanism only offers the capability, but not the guarantee, of producing 3D-aware images, because the supervision of the generator primarily comes from the discriminator. To address this issue, we propose GeoD through learning a geometry-aware discriminator to improve 3D-aware GANs. Concretely, besides differentiating real and fake samples from the 2D image space, the discriminator is additionally asked to derive the geometry information from the inputs, which is then applied as the guidance of the generator. Such a simple yet effective design facilitates learning substantially more accurate 3D shapes. Extensive experiments on various generator architectures and training datasets verify the superiority of GeoD over state-of-the-art alternatives. Moreover, our approach is registered as a general framework such that a more capable discriminator (i.e., with a third task of novel view synthesis beyond domain classification and geometry extraction) can further assist the generator with a better multi-view consistency. Code will be made publicly available.

## [Accelerating Sparse Convolution for Efficient Neural Network Inference](#)

- Yijun Tan · Kai Han · Kang Zhao · Xianzhi Yu · Zidong Du · Yunhe Wang · Jun Yao · Yunji Chen
- abstract@[open-review](#): Weight sparsity is a promising approach to reducing the model size and computation cost of convolutional neural networks (CNNs). Nevertheless, non-zero weights often distribute randomly in sparse CNN models, introducing enormous difficulty in obtaining actual speedup on common hardware (e.g., GPU) over their dense counterparts. Existing acceleration solutions either require hardware modifications for irregular memory access support or rely on a partially structured sparsity pattern. Neither of these methods is capable of achieving fruitful speedup on convolution layers. In this work, we propose an algorithm-software co-designed sparse convolution based on a novel out-vector-wise (OVW) sparse pattern. Building on the insight that vertical vector integrity can preserve continuous memory access in IM2COL, the Ovw pattern treats a \$V \times 1\$ vector as an entirety. To reduce the error caused by sparsity, we propose an equivalent transformation process, i.e., clustering-based channel permutation, to gather similar rows together. Experimental evaluations demonstrate that our method achieves a \$1.7 \times\$ and \$3.2 \times\$ speedup over the SOTA solution and the dense convolution of ResNet50 on NVIDIA V100 at 75% sparsity, respectively, with only negligible accuracy loss. Moreover, compared to the SOTA solution that achieves speedups only on data with 60% sparsity or more, our method begins to obtain speedups on data with only 10% sparsity.

## [Visual Concepts Tokenization](#)

- Tao Yang · Yuwang Wang · Yan Lu · Nanning Zheng
- abstract@[open-review](#): Obtaining the human-like perception ability of abstracting visual concepts from concrete pixels has always been a fundamental and important target in machine learning research fields such as disentangled representation learning and scene decomposition. Towards this goal, we propose an unsupervised transformer-based Visual Concepts Tokenization framework, dubbed VCT, to perceive an image into a set of disentangled visual concept tokens, with each concept token responding to one type of independent visual concept. Particularly, to obtain these concept tokens, we only use cross-attention to extract visual information from the image tokens layer by layer without self-attention between concept tokens, preventing information leakage across concept tokens. We further propose a Concept Disentangling Loss to facilitate that different concept tokens represent independent visual concepts. The cross-attention and disentangling loss play the role of induction and mutual exclusion for the concept tokens, respectively. Extensive experiments on several popular datasets verify the effectiveness of VCT on the tasks of disentangled representation learning and scene decomposition. VCT achieves the state of the art results by a large margin.

## [Segmenting Moving Objects via an Object-Centric Layered Representation](#)

- Junyu Xie · Weidi Xie · Andrew Zisserman
- abstract@[open-review](#): The objective of this paper is a model that is able to discover, track and segment multiple moving objects in a video. We make four contributions: First, we introduce an object-centric segmentation model with a depth-ordered layer representation. This is implemented using a variant of the transformer architecture that ingests optical flow, where each query vector specifies an object and its layer for the entire video. The model can effectively discover multiple moving objects and handle mutual occlusions; Second, we introduce a scalable pipeline for generating multi-object synthetic training data via layer compositions, which is used to train our proposed model, significantly reducing the requirements for labour-intensive annotations, and supporting Sim2Real generalisation; Third, we conduct thorough ablation studies, showing that the model is able to learn object permanence and temporal shape consistency, and is able to predict amodal segmentation masks; Fourth, our synthetic-supervised model is evaluated on standard video segmentation benchmarks, DAVIS, MoCA, SegTrack, FBMS-59, achieving state-of-the-art performance among existing methods that do not rely on any manual annotations. With test-time adaptation, we observe further performance boosts.

## [ACIL: Analytic Class-Incremental Learning with Absolute Memorization and Privacy Protection](#)

- HUIPING ZHUANG · Zhenyu Weng · Hongxin Wei · RENCHUNZI XIE · Kar-Ann Toh · Zhiping Lin
- abstract@[open-review](#): Class-incremental learning (CIL) learns a classification model with training data of different classes arising progressively. Existing CIL either suffers from serious accuracy loss due to catastrophic forgetting, or invades data privacy by revisiting used exemplars. Inspired by learning of linear problems, we propose an analytic class-incremental learning (ACIL) with absolute memorization of past knowledge while avoiding breaching of data privacy (i.e., without storing historical data). The absolute memorization is demonstrated in the sense that the CIL using ACIL given present data would give identical results to that from its joint-learning counterpart that consumes both present and historical samples. This equality is theoretically validated. The data privacy is ensured by showing that no historical data are involved during the learning process. Empirical validations demonstrate ACIL's competitive accuracy performance with near-identical results for various incremental task settings (e.g., 5-50 phases). This also allows ACIL to outperform the state-of-the-art methods for large-phase scenarios (e.g., 25 and 50 phases).

## [MACK: Multimodal Aligned Conceptual Knowledge for Unpaired Image-text Matching](#)

- Yan Huang · Yuming Wang · Yunan Zeng · Liang Wang
- abstract@[open-review](#): Recently, the accuracy of image-text matching has been greatly improved by multimodal pretrained models, all of which are trained on millions or billions of paired images and texts. Different from them, this paper studies a new scenario as unpaired image-text matching, in which paired images and texts are assumed to be unavailable during model training. To deal with this, we propose a simple yet effective method namely Multimodal Aligned Conceptual Knowledge (MACK), which is inspired by the knowledge use in human brain. It can be directly used as general knowledge to correlate images and texts even without model training, or further fine-tuned based on unpaired images and texts to better generalize to certain datasets. In addition, we extend it as a re-ranking method, which can be easily combined with existing image-text matching models to substantially improve their performance.

## [Real-Valued Backpropagation is Unsuitable for Complex-Valued Neural Networks](#)

- ZHIHAO TAN · Yi Xie · Yuan Jiang · Zhi-Hua Zhou
- abstract@[open-review](#): Recently complex-valued neural networks have received increasing attention due to successful applications in various tasks and the potential advantages of better theoretical properties and richer representational capacity. However, the training dynamics of complex networks compared to real networks remains an open problem. In this paper, we investigate the dynamics of deep complex networks during real-valued

backpropagation in the infinite-width limit via neural tangent kernel (NTK). We first extend the Tensor Program to the complex domain, to show that the dynamics of any basic complex network architecture is governed by its NTK under real-valued backpropagation. Then we propose a way to investigate the comparison of training dynamics between complex and real networks by studying their NTKs. As a result, we surprisingly prove that for most complex activation functions, the commonly used real-valued backpropagation reduces the training dynamics of complex networks to that of ordinary real networks, thus eliminating the characteristics of complex-valued neural networks. Finally, we study the results numerically and the experiments verify our theoretical findings.

## [Learning Latent Seasonal-Trend Representations for Time Series Forecasting](#)

- Zhiyuan Wang · Xovee Xu · Goce Trajcevski · Weifeng Zhang · Ting Zhong · Fan Zhou
- abstract@[open-review](#): Forecasting complex time series is ubiquitous and vital in a range of applications but challenging. Recent advances endeavor to achieve progress by incorporating various deep learning techniques (e.g., RNN and Transformer) into sequential models. However, clear patterns are still hard to extract since intricate time series are composed of several entangled components. Motivated by the success of disentangled variational autoencoder in computer vision and classical time series decomposition, we plan to infer a couple of representations that depict seasonal and trend components of time series. To achieve this goal, we propose LaST, which, based on variational inference, aims to disentangle the seasonal-trend representations in the latent space. Furthermore, LaST supervises and disassociates representations from the perspectives of themselves and input reconstruction, and introduces a series of auxiliary objectives. Extensive experiments prove that LaST achieves state-of-the-art performance on time series forecasting task with 23.7% and 20.5% relative improvement against the most advanced representation learning and end-to-end forecasting models. For reproducibility, our source code has been submitted in the supplementary material.

## [Semi-Discrete Normalizing Flows through Differentiable Tessellation](#)

- Ricky T. Q. Chen · Brandon Amos · Maximilian Nickel
- abstract@[open-review](#): Mapping between discrete and continuous distributions is a difficult task and many have had to resort to heuristical approaches. We propose a tessellation-based approach that directly learns quantization boundaries in a continuous space, complete with exact likelihood evaluations. This is done through constructing normalizing flows on convex polytopes parameterized using a simple homeomorphism with an efficient log determinant Jacobian. We explore this approach in two application settings, mapping from discrete to continuous and vice versa. Firstly, a Voronoi dequantization allows automatically learning quantization boundaries in a multidimensional space. The location of boundaries and distances between regions can encode useful structural relations between the quantized discrete values. Secondly, a Voronoi mixture model has constant computation cost for likelihood evaluation regardless of the number of mixture components. Empirically, we show improvements over existing methods across a range of structured data modalities.

## [Optimal-er Auctions through Attention](#)

- Dmitry Ivanov · Iskander Saifulin · Igor Filippov · Ksenia Balabaeva
- abstract@[open-review](#): RegretNet is a recent breakthrough in the automated design of revenue-maximizing auctions. It combines the expressivity of deep learning with the regret-based approach to relax the Incentive Compatibility constraint (that participants benefit from bidding truthfully). We propose two independent modifications of RegretNet, namely a new neural architecture based on the attention mechanism, denoted as RegretFormer, and a new interpretable loss function that is significantly less sensitive to hyperparameters. We investigate both proposed modifications in an extensive experimental study and additionally test out-of-setting generalization of our network. In all experiments, we find that RegretFormer consistently outperforms existing architectures in revenue. Regarding our loss modification, we confirm its effectiveness in controlling the revenue-regret trade-off by varying a single interpretable hyperparameter.

## [Minimax-Optimal Multi-Agent RL in Zero-Sum Markov Games With a Generative Model](#)

- Gen Li · Yuejie Chi · Yuxin Chen · Yuting Wei
- abstract@[open-review](#): This paper is concerned with two-player zero-sum Markov games --- arguably the most basic setting in multi-agent reinforcement learning --- with the goal of learning a Nash equilibrium (NE) sample-optimally. All prior results suffer from at least one of the two obstacles: the curse of multiple agents and the barrier of long horizon, regardless of the sampling protocol in use. We take a first step towards settling this problem, assuming access to a flexible sampling mechanism: the generative model. Focusing on non-stationary finite-horizon Markov games, we develop a learning algorithm and an adaptive sampling scheme that leverage the optimism principle in adversarial learning (particularly the Follow-the-Regularized-Leader (FTRL) method), with a delicate design of bonus terms that ensure certain decomposability under the FTRL dynamics. Our algorithm learns an  $\$varepsilon$ -approximate Markov NE policy using  $\$widetilde{O}(\frac{H^4 S(A+B)}{\$varepsilon^2})\$samples$ , where  $S$  is the number of states,  $H$  is the horizon, and  $A$  (resp.  $B$ ) is the number of actions for the max-player (resp. min-player). This is nearly un-improvable in a minimax sense. Along the way, we derive a refined regret bound for FTRL that makes explicit the role of variance-type quantities.

## [Sharpness-Aware Training for Free](#)

- JIAWEI DU · Daquan Zhou · Joey Tianyi Zhou · Jiashi Feng · Vincent Tan
- abstract@[open-review](#): Modern deep neural networks (DNNs) have achieved state-of-the-art performances but are typically over-parameterized. The over-parameterization may result in undesirably large generalization error in the absence of other customized training strategies. Recently, a line of research under the name of Sharpness-Aware Minimization (SAM) has shown that minimizing a sharpness measure, which reflects the geometry of the loss landscape, can significantly reduce the generalization error. However, SAM-like methods incur a two-fold computational overhead of the given base optimizer (e.g. SGD) for approximating the sharpness measure. In this paper, we propose Sharpness-Aware Training for Free, or SAF, which mitigates the sharp landscape at almost zero additional computational cost over the base optimizer. Intuitively, SAF achieves this by avoiding sudden drops in the loss in the sharp local minima throughout the trajectory of the updates of the weights. Specifically, we suggest a novel trajectory loss, based on the KL-divergence between the outputs of DNNs with the current weights and past weights, as a replacement of the SAM's sharpness measure. This loss captures the rate of change of the training loss along the model's update trajectory. By minimizing it, SAF ensures the convergence to a flat minimum with improved generalization capabilities. Extensive empirical results show that SAF minimizes the sharpness in the same way that SAM does, yielding better results on the ImageNet dataset with essentially the same computational cost as the base optimizer.

## [Efficient and Effective Optimal Transport-Based Biclustering](#)

- Chakib Fettal · Iazhar labiod · Mohamed NADIF
- abstract@[open-review](#): Bipartite graphs can be used to model a wide variety of dyadic information such as user-rating, document-term, and gene-disorder pairs. Biclustering is an extension of clustering to the underlying bipartite graph induced from this kind of data. In this paper, we leverage optimal transport (OT) which has gained momentum in the machine learning community to propose a novel and scalable biclustering model that generalizes several classical biclustering approaches. We perform extensive experimentation to show the validity of our approach compared to other OT biclustering algorithms along both dimensions of the dyadic datasets.

## [Beyond Mahalanobis Distance for Textual OOD Detection](#)

- Pierre Colombo · Eduardo Dadalto · Guillaume Staerman · Nathan Noiry · Pablo Piantanida
- abstract@[open-review](#): As the number of AI systems keeps growing, it is fundamental to implement and develop efficient control mechanisms to ensure the safe and proper functioning of machine learning (ML) systems. Reliable out-of-distribution (OOD) detection aims to detect test samples that are statistically far from the training distribution, as they might cause failures of in-production systems. In this paper, we propose a new detector called TRUSTED. Different from previous works, TRUSTED key components (i) include a novel OOD score relying on the concept of statistical data depth, (ii) rely on the idea's full potential that all hidden layers of the network carry information regarding OOD. Our extensive experiments, comparing over 51k model configurations including different checkpoints, seed and various datasets, demonstrate that TRUSTED achieve state-of-the-art performances by producing an improvement of over 3 AUROC points.

## [Augmented RBMLE-UCB Approach for Adaptive Control of Linear Quadratic Systems](#)

- Akshay Mete · Rahul Singh · P. R. Kumar
- abstract@[open-review](#): We consider the problem of controlling a stochastic linear system with quadratic costs, when its system parameters are not known to the agent -- called the adaptive LQ control problem. We re-examine an approach called Reward-Biased Maximum Likelihood Estimate'' (RBMLE) that was proposed more than forty years ago, and which predates the "Upper Confidence Bound" (UCB) method as well as the definition of ``regret''. It simply added a term favoring parameters with larger rewards to the criterion for parameter estimation. We show how the RBMLE and UCB methods can be reconciled, and thereby propose an Augmented RBMLE-UCB algorithm that combines the penalty of the RBMLE method with the constraint of the UCB method, uniting the two approaches to optimism in the face of uncertainty. We establish that theoretically this method retains  $\mathcal{O}(\sqrt{T})$  regret, the best known so far. We further compare the empirical performance of the proposed Augmented RBMLE-UCB and the standard RBMLE (without the augmentation) with UCB (OFU), Thompson Sampling (TS), Input Perturbation (IP), Randomized Certainty Equivalence (RCE) and StabL on many real-world examples that include flight control of Boeing 747 and Unmanned Aerial Vehicle. We perform extensive simulation studies showing that the Augmented RBMLE consistently outperforms OFU, TS and StabL by a huge margin, while it is marginally better than IP and moderately better than RCE.

## [MetricFormer: A Unified Perspective of Correlation Exploring in Similarity Learning](#)

- Jieyi Yan · Erkun Yang · Cheng Deng · Heng Huang
- abstract@[open-review](#): Similarity learning can be significantly advanced by informative relationships among different samples and features. The current methods try to excavate the multiple correlations in different aspects, but cannot integrate them into a unified framework. In this paper, we provide to consider the multiple correlations from a unified perspective and propose a new method called MetricFormer, which can effectively capture and model the multiple correlations with an elaborate metric transformer. In MetricFormer, the feature decoupling block is adopted to learn an ensemble of distinct and diverse features with different discriminative characteristics. After that, we apply the batch-wise correlation block into the batch dimension of each mini-batch to implicitly explore sample relationships. Finally, the feature-wise correlation block is performed to discover the intrinsic structural pattern of the ensemble of features and obtain the aggregated feature embedding for similarity measuring. With three kinds of transformer blocks, we can learn more representative features through the proposed MetricFormer. Moreover, our proposed method can be flexibly integrated with any metric learning framework. Extensive experiments on three widely-used datasets demonstrate the superiority of our proposed method over state-of-the-art methods.

## [A Unified Diversity Measure for Multiagent Reinforcement Learning](#)

- Zongkai Liu · Chao Yu · Yaodong Yang · peng sun · Zifan Wu · Yuan Li
- abstract@[open-review](#): Promoting behavioural diversity is of critical importance in multi-agent reinforcement learning, since it helps the agent population maintain robust performance when encountering unfamiliar opponents at test time, or, when the game is highly non-transitive in the strategy space (e.g., Rock-Paper-Scissor). While a myriad of diversity metrics have been proposed, there are no widely accepted or unified definitions in the literature, making the consequent diversity-aware learning algorithms difficult to evaluate and the insights elusive. In this work, we propose a novel metric called the Unified Diversity Measure (UDM) that offers a unified view for existing diversity metrics. Based on UDM, we design the UDM-Fictitious Play (UDM-FP) and UDM-Policy Space Response Oracle (UDM-PSRO) algorithms as efficient solvers for normal-form games and open-ended games. In theory, we prove that UDM-based methods can enlarge the gamescape by increasing the response capacity of the strategy pool, and have convergence guarantee to two-player Nash equilibrium. We validate our algorithms on games that show strong non-transitivity, and empirical results show that our algorithms achieve better performances than strong PSRO baselines in terms of the exploitability and population effectivity.

## [Dynamic Pricing with Monotonicity Constraint under Unknown Parametric Demand Model](#)

- Su Jia · Andrew Li · R Ravi
- abstract@[open-review](#): We consider the Continuum Bandit problem where the goal is to find the optimal action under an unknown reward function, with an additional monotonicity constraint (or, "markdown" constraint) that requires that the action sequence be non-increasing. This problem faithfully models a natural single-product dynamic pricing problem, called "markdown pricing", where the objective is to adaptively reduce the price over a finite sales horizon to maximize expected revenues. Jia et al '21 and Chen '21 independently showed a tight  $T^{3/4}$  regret bound over  $T$  rounds under *minimal* assumptions of unimodality and Lipschitzness in the reward (or, "revenue") function. This bound shows that the demand learning in markdown pricing is harder than unconstrained (i.e., without the monotonicity constraint) pricing under unknown demand which suffers regret only of the order of  $T^{2/3}$  under the same assumptions (Kleinberg '04). However, in practice the demand functions are usually assumed to have certain functional forms (e.g. linear or exponential), rendering the demand-learning easier and suggesting lower regret bounds. We investigate two fundamental questions, assuming the underlying demand curve comes from a given parametric family: (1) Can we improve the  $T^{3/4}$  regret bound for markdown pricing, under extra assumptions on the functional forms of the demand functions? (2) Is markdown pricing still harder than unconstrained pricing, under these additional assumptions? To answer these, we introduce a concept called markdown dimension that measures the complexity of the parametric family and present tight regret bounds under this framework, thereby completely settling the aforementioned questions.

## [Near-Optimal Private and Scalable \$k\$ -Clustering](#)

- Vincent Cohen-Addad · Alessandro Epasto · Vahab Mirrokni · Shyam Narayanan · Peilin Zhong
- abstract@[open-review](#): We study the differentially private (DP)  $k$ -means and  $k$ -median clustering problems of  $n$  points in  $d$ -dimensional Euclidean space in the massively parallel computation (MPC) model. We provide two near-optimal algorithms where the near-optimality is in three aspects: they both achieve (1).  $O(1)$  parallel computation rounds, (2). near-linear in  $n$  and polynomial in  $k$  total computational work (i.e., near-linear running time in the sequential setting), (3).  $O(1)$  relative approximation and  $\text{poly}(k, d)$  additive error, where  $\Omega(1)$  relative approximation is provably necessary even for any polynomial-time non-private algorithm, and  $\Omega(k)$  additive error is a provable lower bound for any polynomial-time DP  $k$ -means/median algorithm. Our two algorithms provide a tradeoff between the relative approximation and the additive error: the first has  $\sim (k^{2.5} + k^{1.01}) \sqrt{d}$  additive error, and the second one achieves  $(1+\gamma)$  relative approximation to the optimal non-private algorithm for an arbitrary small constant  $\gamma > 0$  and with  $\text{poly}(k, d)$  additive error for a larger polynomial dependence on  $k$  and  $d$ . To achieve our result, we develop a general framework which partitions the data and reduces the DP clustering problem for the entire dataset to the DP clustering problem for each part. To control the blow-up of the additive error introduced by each part, we develop a novel charging argument which might be of independent interest.

## [Branch & Learn for Recursively and Iteratively Solvable Problems in Predict+Optimize](#)

- Xinyi Hu · Jasper Lee · Jimmy Lee · Allen Z. Zhong
- abstract@[open-review](#): This paper proposes Branch & Learn, a framework for Predict+Optimize to tackle optimization problems containing parameters that are unknown at the time of solving. Given an optimization problem solvable by a recursive algorithm satisfying simple conditions, we show how a corresponding learning algorithm can be constructed directly and methodically from the recursive algorithm. Our framework applies also to iterative algorithms by viewing them as a degenerate form of recursion. Extensive experimentation shows better performance for our proposal over classical and state of the art approaches.

## [Generalizing Bayesian Optimization with Decision-theoretic Entropies](#)

- Willie Neiswanger · Lantao Yu · Shengjia Zhao · Chenlin Meng · Stefano Ermon
- abstract@[open-review](#): Bayesian optimization (BO) is a popular method for efficiently inferring optima of an expensive black-box function via a sequence of queries. Existing information-theoretic BO procedures aim to make queries that most reduce the uncertainty about optima, where the uncertainty is captured by Shannon entropy. However, an optimal measure of uncertainty would, ideally, factor in how we intend to use the inferred quantity in some downstream procedure. In this paper, we instead consider a generalization of Shannon entropy from work in statistical decision theory (DeGroot 1962, Rao 1984), which contains a broad class of uncertainty measures parameterized by a problem-specific loss function corresponding to a downstream task. We first show that special cases of this entropy lead to popular acquisition functions used in BO procedures such as knowledge gradient, expected improvement, and entropy search. We then show how alternative choices for the loss yield a flexible family of acquisition functions that can be customized for use in novel optimization settings. Additionally, we develop gradient-based methods to efficiently optimize our proposed family of acquisition functions, and demonstrate strong empirical performance on a diverse set of sequential decision making tasks, including variants of top-\$k\$ optimization, multi-level set estimation, and sequence search.

## [A Non-Asymptotic Moreau Envelope Theory for High-Dimensional Generalized Linear Models](#)

- Lijia Zhou · Frederic Koehler · Pragya Sur · Danica J. Sutherland · Nati Srebro
- abstract@[open-review](#): We prove a new generalization bound that shows for any class of linear predictors in Gaussian space, the Rademacher complexity of the class and the training error under any continuous loss  $\|\cdot\|$  can control the test error under all Moreau envelopes of the loss  $\|\cdot\|$ . We use our finite-sample bound to directly recover the “optimistic rate” of Zhou et al. (2021) for linear regression with the square loss, which is known to be tight for minimal  $\|\cdot\|_2$ -norm interpolation, but we also handle more general settings where the label is generated by a potentially misspecified multi-index model. The same argument can analyze noisy interpolation of max-margin classifiers through the squared hinge loss, and establishes consistency results in spiked-covariance settings. More generally, when the loss is only assumed to be Lipschitz, our bound effectively improves Talagrand’s well-known contraction lemma by a factor of two, and we prove uniform convergence of interpolators (Koehler et al. 2021) for all smooth, non-negative losses. Finally, we show that application of our generalization bound using localized Gaussian width will generally be sharp for empirical risk minimizers, establishing a non-asymptotic Moreau envelope theory for generalization that applies outside of proportional scaling regimes, handles model misspecification, and complements existing asymptotic Moreau envelope theories for M-estimation.

## [Statistically Meaningful Approximation: a Case Study on Approximating Turing Machines with Transformers](#)

- Colin Wei · Yining Chen · Tengyu Ma
- abstract@[open-review](#): A common lens to theoretically study neural net architectures is to analyze the functions they can approximate. However, the constructions from approximation theory often have unrealistic aspects, for example, reliance on infinite precision to memorize target function values. To address this issue, we propose a formal definition of statistically meaningful approximation which requires the approximating network to exhibit good statistical learnability. We present case studies on statistically meaningful approximation for two classes of functions: boolean circuits and Turing machines. We show that overparameterized feedforward neural nets can statistically meaningfully approximate boolean circuits with sample complexity depending only polynomially on the circuit size, not the size of the approximating network. In addition, we show that transformers can statistically meaningfully approximate Turing machines with computation time bounded by  $T$ , requiring sample complexity polynomial in the alphabet size, state space size, and  $\log(T)$ . Our analysis introduces new tools for generalization bounds that provide much tighter sample complexity guarantees than the typical VC-dimension or norm-based bounds, which may be of independent interest.

## [On Learning Fairness and Accuracy on Multiple Subgroups](#)

- Changjian Shui · Gezheng Xu · Qi CHEN · Jiaqi Li · Charles Ling · Tal Arbel · Boyu Wang · Christian Gagné
- abstract@[open-review](#): We propose an analysis in fair learning that preserves the utility of the data while reducing prediction disparities under the criteria of group sufficiency. We focus on the scenario where the data contains multiple or even many subgroups, each with limited number of samples. As a result, we present a principled method for learning a fair predictor for all subgroups via formulating it as a bilevel objective. Specifically, the subgroup specific predictors are learned in the lower-level through a small amount of data and the fair predictor. In the upper-level, the fair predictor is updated to be close to all subgroup specific predictors. We further prove that such a bilevel objective can effectively control the group sufficiency and generalization error. We evaluate the proposed framework on real-world datasets. Empirical evidence suggests the consistently improved fair predictions, as well as the comparable accuracy to the baselines.

## [Locally Hierarchical Auto-Regressive Modeling for Image Generation](#)

- Tackgeun You · Saehoon Kim · Chiheon Kim · Doyup Lee · Bohyung Han
- abstract@[open-review](#): Auto-Regressive (AR) modeling on image generation provides a scalable solution to synthesize high-quality examples by directly optimizing the data likelihood, factorized into a product of conditionals in pre-defined generation order. However, it is computationally expensive to train AR models on the pixels of high-resolution images, and many practical approaches decompose a learning process into the following two stages: (a) representing a high-resolution image by a short discrete sequence, and (b) learning an AR model on this sequence, rather than on pixels. This paper points out that every code in the sequence has a receptive field of the same size and most existing works in this framework fail to exploit multi-scale information of images. To mitigate this limitation, we represent an image by a set of discrete sequences of multiple resolutions and generate images using an AR model based on variational autoencoder, which is referred to as Hierarchical-Quantized Variational AutoEncoder (HQ-VAE). We propose a Hierarchical-Quantized Transformers (HQ-Transformer) to model the multi-level discrete sequences efficiently and generate novel images of good quality. Experiments on class-conditional and text-conditional generation tasks verify that our approach outperforms previous AR models, in terms of fidelity of generation samples under a similar computation budget. We also demonstrate structure-condition image generation to show the usage of disentangled top and bottom codes. Extensive ablation studies explain rationales for our architectural design.

## [A Rotated Hyperbolic Wrapped Normal Distribution for Hierarchical Representation Learning](#)

- Seunghyuk Cho · Juyong Lee · Jaesik Park · Dongwoo Kim
- abstract@[open-review](#): We present a rotated hyperbolic wrapped normal distribution (RoWN), a simple yet effective alteration of a hyperbolic wrapped normal distribution (HWN). The HWN expands the domain of probabilistic modeling from Euclidean to hyperbolic space, where a tree can be embedded with arbitrary low distortion in theory. In this work, we analyze the geometric properties of the diagonal HWN, a standard choice of distribution in probabilistic modeling. The analysis shows that the distribution is inappropriate to represent the data points at the same hierarchy level through their angular distance with the same norm in the Poincaré disk model. We then empirically verify the presence of limitations of HWN, and show how RoWN,

the proposed distribution, can alleviate the limitations on various hierarchical datasets, including noisy synthetic binary tree, WordNet, and Atari 2600 Breakout.

## [Learning Generalized Policy Automata for Relational Stochastic Shortest Path Problems](#)

- Rushang Karia · Rashmeet Kaur Nayyar · Siddharth Srivastava
- abstract@[open-review](#): Several goal-oriented problems in the real-world can be naturally expressed as Stochastic Shortest Path Problems (SSPs). However, the computational complexity of solving SSPs makes finding solutions to even moderately sized problems intractable. Currently, existing state-of-the-art planners and heuristics often fail to exploit knowledge learned from solving other instances. This paper presents an approach for learning \emph{Generalized Policy Automata} (GPA): non-deterministic partial policies that can be used to catalyze the solution process. GPAs are learned using relational, feature-based abstractions, which makes them applicable on broad classes of related problems with different object names and quantities. Theoretical analysis of this approach shows that it guarantees completeness and hierarchical optimality. Empirical analysis shows that this approach effectively learns broadly applicable policy knowledge in a few-shot fashion and significantly outperforms state-of-the-art SSP solvers on test problems whose object counts are far greater than those used during training.

## [Towards Consistency in Adversarial Classification](#)

- Laurent Meunier · Raphael Etedgui · Rafael Pinot · Yann Chevaleyre · Jamal Atif
- abstract@[open-review](#): In this paper, we study the problem of consistency in the context of adversarial examples. Specifically, we tackle the following question: can surrogate losses still be used as a proxy for minimizing the \$0/1\$ loss in the presence of an adversary that alters the inputs at test-time? Different from the standard classification task, this question cannot be reduced to a point-wise minimization problem, and calibration needs not to be sufficient to ensure consistency. In this paper, we expose some pathological behaviors specific to the adversarial problem, and show that no convex surrogate loss can be consistent or calibrated in this context. It is therefore necessary to design another class of surrogate functions that can be used to solve the adversarial consistency issue. As a first step towards designing such a class, we identify sufficient and necessary conditions for a surrogate loss to be calibrated in both the adversarial and standard settings. Finally, we give some directions for building a class of losses that could be consistent in the adversarial framework.

## [Distilled Gradient Aggregation: Purify Features for Input Attribution in the Deep Neural Network](#)

- Giyoung Jeon · Haedong Jeong · Jaesik Choi
- abstract@[open-review](#): Measuring the attribution of input features toward the model output is one of the popular post-hoc explanations on the Deep Neural Networks (DNNs). Among various approaches to compute the attribution, the gradient-based methods are widely used to generate attributions, because of its ease of implementation and the model-agnostic characteristic. However, existing gradient integration methods such as Integrated Gradients (IG) suffer from (1) the noisy attributions which cause the unreliability of the explanation, and (2) the selection for the integration path which determines the quality of explanations. FullGrad (FG) is an another approach to construct the reliable attributions by focusing the locality of piece-wise linear network with the bias gradient. Although FG has shown reasonable performance for the given input, as the shortage of the global property, FG is vulnerable to the small perturbation, while IG which includes the exploration over the input space is robust. In this work, we design a new input attribution method which adopt the strengths of both local and global attributions. In particular, we propose a novel approach to distill input features using weak and extremely positive contributor masks. We aggregate the intermediate local attributions obtained from the distillation sequence to provide reliable attribution. We perform the quantitative evaluation compared to various attribution methods and show that our method outperforms others. We also provide the qualitative result that our method obtains object-aligned and sharp attribution heatmap.

## [Optimal Positive Generation via Latent Transformation for Contrastive Learning](#)

- Yinqi Li · Hong Chang · Bingpeng MA · Shiguang Shan · Xilin Chen
- abstract@[open-review](#): Contrastive learning, which learns to contrast positive with negative pairs of samples, has been popular for self-supervised visual representation learning. Although great effort has been made to design proper positive pairs through data augmentation, few works attempt to generate optimal positives for each instance. Inspired by semantic consistency and computational advantage in latent space of pretrained generative models, this paper proposes to learn instance-specific latent transformations to generate Contrastive Optimal Positives (COP-Gen) for self-supervised contrastive learning. Specifically, we formulate COP-Gen as an instance-specific latent space navigator which minimizes the mutual information between the generated positive pair subject to the semantic consistency constraint. Theoretically, the learned latent transformation creates optimal positives for contrastive learning, which removes as much nuisance information as possible while preserving the semantics. Empirically, using generated positives by COP-Gen consistently outperforms other latent transformation methods and even real-image-based methods in self-supervised contrastive learning.

## [Polynomial time guarantees for the Burer-Monteiro method](#)

- Diego Cifuentes · Ankur Moitra
- abstract@[open-review](#): The Burer-Monteiro method is one of the most widely used techniques for solving large-scale semidefinite programs (SDP). The basic idea is to solve a nonconvex program in  $\mathbb{Y}$ , where  $\mathbb{Y}$  is an  $n \times p$  matrix such that  $\mathbb{X} = \mathbb{Y} \mathbb{Y}^T$ . We show that this method can solve SDPs in polynomial time in a smoothed analysis setting. More precisely, we consider an SDP whose domain satisfies some compactness and smoothness assumptions, and slightly perturb the cost matrix and the constraints. We show that if  $p \geq \sqrt{2(1+\eta)m}$ , where  $m$  is the number of constraints and  $\eta > 0$  is any fixed constant, then the Burer-Monteiro method can solve SDPs to any desired accuracy in polynomial time, in the setting of smooth analysis. The bound on  $p$  approaches the celebrated Barylinok-Pataki bound in the limit as  $\eta$  goes to zero, beneath which the nonconvex program can be suboptimal. Our main technical contribution, which is key for our tight bound on  $p$ , is to connect spurious approximately critical points of the nonconvex program to tubular neighborhoods of certain algebraic varieties, and then estimate the volume of such tubes.

## [Whitening Convergence Rate of Affine Coupling Flows](#)

- Felix Draxler · Christoph Schnorr · Ullrich Käthe
- abstract@[open-review](#): Coupling flows (e.g. RealNVP) are a popular family of normalizing flow architectures that work surprisingly well in practice. This calls for theoretical understanding. Existing work shows that such flows weakly converge to arbitrary data distributions. However, they make no statement about the stricter convergence criterion used in practice, the maximum likelihood loss. For the first time, we make a quantitative statement about this kind of convergence: We prove that coupling flows perform whitening of the data distribution (i.e. diagonalize the covariance matrix) and derive corresponding convergence bounds that show a linear convergence rate in the depth of the flow. Numerical experiments demonstrate the implications of our theory and point at open questions.

## [Towards Improving Calibration in Object Detection Under Domain Shift](#)

- Muhammad Akhtar Munir · Muhammad Haris Khan · M. Sarfraz · Mohsen Ali
- abstract@[open-review](#): The increasing use of deep neural networks in safety-critical applications requires the trained models to be well-calibrated. Most current calibration techniques address classification problems while focusing on improving calibration on in-domain predictions. Little to no attention is

paid towards addressing calibration of visual object detectors which occupy similar space and importance in many decision making systems. In this paper, we study the calibration of current object detection models, particularly under domain shift. To this end, we first introduce a plug-and-play train-time calibration loss for object detection. It can be used as an auxiliary loss function to improve detector's calibration. Second, we devise a new uncertainty quantification mechanism for object detection which can implicitly calibrate the commonly used self-training based domain adaptive detectors. We include in our study both single-stage and two-stage object detectors. We demonstrate that our loss improves calibration for both in-domain and out-of-domain detections with notable margins. Finally, we show the utility of our techniques in calibrating the domain adaptive object detectors in diverse domain shift scenarios.

## [ZSON: Zero-Shot Object-Goal Navigation using Multimodal Goal Embeddings](#)

- Arjun Majumdar · Gunjan Aggarwal · Bhavika Devnani · Judy Hoffman · Dhruv Batra
- abstract@[open-review](#): We present a scalable approach for learning open-world object-goal navigation (ObjectNav). Our approach is zero-shot -- i.e., it does not require ObjectNav annotations for training. Instead, we convert the image-goal navigation task into semantic-goal navigation (SemanticNav) by encoding goal images into a multimodal, semantic embedding space. SemanticNav agents can be trained at scale in unannotated 3D environments (e.g., HM3D). After training, the resulting agents can navigate to objects described in free-form natural language (e.g., ‘sink’, ‘bathroom sink’, etc.), reflecting the requirements for open-world ObjectNav. We extensively evaluate our agents on three ObjectNav datasets (Gibson, HM3D, and MP3D) and observe absolute improvements in success of 4.2% - 20.0% over existing zero-shot methods. In an open-world setting, we discover that our agents can generalize to compound instructions with a room explicitly mentioned (e.g., ‘Find a kitchen sink’) and when the target room can be inferred (e.g., ‘Find a sink and a stove’).

## [An Analysis of Ensemble Sampling](#)

- Chao Qin · Zheng Wen · Xiuyuan Lu · Benjamin Van Roy
- abstract@[open-review](#): Ensemble sampling serves as a practical approximation to Thompson sampling when maintaining an exact posterior distribution over model parameters is computationally intractable. In this paper, we establish a regret bound that ensures desirable behavior when ensemble sampling is applied to the linear bandit problem. This represents the first rigorous regret analysis of ensemble sampling and is made possible by leveraging information-theoretic concepts and novel analytic techniques that may prove useful beyond the scope of this paper.

## [M2N: Mesh Movement Networks for PDE Solvers](#)

- Wenbin Song · Mingrui Zhang · Joseph G Wallwork · Junpeng Gao · Zheng Tian · Fanglei Sun · Matthew Piggott · Junqing Chen · Zuoqiang Shi · Xiang Chen · Jun Wang
- abstract@[open-review](#): Numerical Partial Differential Equation (PDE) solvers often require discretizing the physical domain by using a mesh. Mesh movement methods provide the capability to improve the accuracy of the numerical solution without introducing extra computational burden to the PDE solver, by increasing mesh resolution where the solution is not well-resolved, whilst reducing unnecessary resolution elsewhere. However, sophisticated mesh movement methods, such as the Monge-Ampère method, generally require the solution of auxiliary equations. These solutions can be extremely expensive to compute when the mesh needs to be adapted frequently. In this paper, we propose to the best of our knowledge the first learning-based end-to-end mesh movement framework for PDE solvers. Key requirements of learning-based mesh movement methods are: alleviating mesh tangling, boundary consistency, and generalization to mesh with different resolutions. To achieve these goals, we introduce the neural spline model and the graph attention network (GAT) into our models respectively. While the Neural-Spline based model provides more flexibility for large mesh deformation, the GAT based model can handle domains with more complicated shapes and is better at performing delicate local deformation. We validate our methods on stationary and time-dependent, linear and non-linear equations, as well as regularly and irregularly shaped domains. Compared to the traditional Monge-Ampère method, our approach can greatly accelerate the mesh adaptation process by three to four orders of magnitude, whilst achieving comparable numerical error reduction.

## [On the symmetries of the synchronization problem in Cryo-EM: Multi-Frequency Vector Diffusion Maps on the Projective Plane](#)

- Gabriele Cesa · Arash Behboodi · Taco Cohen · Max Welling
- abstract@[open-review](#): Cryo-Electron Microscopy (Cryo-EM) is an important imaging method which allows high-resolution reconstruction of the 3D structures of biomolecules. It produces highly noisy 2D images by projecting a molecule's 3D density from random viewing directions. Because the projection directions are unknown, estimating the images' poses is necessary to perform the reconstruction. We focus on this task and study it under the group synchronization framework: if the relative poses of pairs of images can be approximated from the data, an estimation of the images' poses is given by the assignment which is most consistent with the relative ones. In particular, by studying the symmetries of cryo-EM, we show that relative poses in the group O(2) provide sufficient constraints to identify the images' poses, up to the molecule's chirality. With this in mind, we improve the existing multi-frequency vector diffusion maps (MFVDM) method: by using O(2) relative poses, our method not only predicts the similarity between the images' viewing directions but also recovers their poses. Hence, we can leverage all input images in a 3D reconstruction algorithm by initializing the poses with our estimation rather than just clustering and averaging the input images. We validate the recovery capabilities and robustness of our method on randomly generated synchronization graphs and a synthetic cryo-EM dataset.

## [What is Where by Looking: Weakly-Supervised Open-World Phrase-Grounding without Text Inputs](#)

- Tal Shaharabany · Yoad Tewel · Lior Wolf
- abstract@[open-review](#): Given an input image, and nothing else, our method returns the bounding boxes of objects in the image and phrases that describe the objects. This is achieved within an open world paradigm, in which the objects in the input image may not have been encountered during the training of the localization mechanism. Moreover, training takes place in a weakly supervised setting, where no bounding boxes are provided. To achieve this, our method combines two pre-trained networks: the CLIP image-to-text matching score and the BLIP image captioning tool. Training takes place on COCO images and their captions and is based on CLIP. Then, during inference, BLIP is used to generate a hypothesis regarding various regions of the current image. Our work generalizes weakly supervised segmentation and phrase grounding and is shown empirically to outperform the state of the art in both domains. It also shows very convincing results in the novel task of weakly-supervised open-world purely visual phrase-grounding presented in our work. For example, on the datasets used for benchmarking phrase-grounding, our method results in a very modest degradation in comparison to methods that employ human captions as an additional input.

## [Algorithms and Hardness for Learning Linear Thresholds from Label Proportions](#)

- Rishi Saket
- abstract@[open-review](#): We study the learnability of linear threshold functions (LTFs) in the learning from label proportions (LLP) framework. In this, the feature-vector classifier is learnt from bags of feature-vectors and their corresponding observed label proportions which are satisfied by (i.e., consistent with) some unknown LTF. This problem has been investigated in recent work (Saket21) which gave an algorithm to produce an LTF that satisfies at least \$(2/5)\$-fraction of a satisfiable collection of bags, each of size \$\leq 2\$, by solving and rounding a natural SDP relaxation. However, this SDP relaxation is specific to at most \$2\$-sized bags and does not apply to bags of larger size. In this work we provide a fairly non-trivial SDP relaxation of a non-quadratic formulation for bags of size \$3\$. We analyze its rounding procedure using novel matrix decomposition techniques to obtain an algorithm which

outputs an LTF satisfying at least  $(1/12)$ -fraction of the bags of size  $\leq 3$ . We also apply our techniques to bags of size  $\geq 4$  to provide a  $\Omega(1/q)$ -approximation guarantee for a weaker notion of satisfiability. We include comparative experiments on simulated data demonstrating the applicability of our algorithmic techniques. From the complexity side we provide a hardness reduction to produce instances with bags of any constant size  $q$ . Our reduction proves the NP-hardness of satisfying more than  $\lceil \frac{1}{q} \rceil + o(1)$  fraction of a satisfiable collection of such bags using as hypothesis any function of constantly many LTFs, showing thereby that the problem is harder to approximate as the bag size  $q$  increases. Using a strengthened analysis, for  $q=2$  we obtain a  $\lceil \frac{4}{9} \rceil + o(1)$  hardness factor for this problem, improving upon the  $\lceil \frac{1}{2} \rceil + o(1)$  factor shown by Saket21.

## [RecursiveMix: Mixed Learning with History](#)

- Lingfeng Yang · Xiang Li · Borui Zhao · Renjie Song · Jian Yang
- abstract@[open-review](#): Mix-based augmentation has been proven fundamental to the generalization of deep vision models. However, current augmentations only mix samples from the current data batch during training, which ignores the possible knowledge accumulated in the learning history. In this paper, we propose a recursive mixed-sample learning paradigm, termed “RecursiveMix” (RM), by exploring a novel training strategy that leverages the historical input-prediction-label triplets. More specifically, we iteratively resize the input image batch from the previous iteration and paste it into the current batch while their labels are fused proportionally to the area of the operated patches. Furthermore, a consistency loss is introduced to align the identical image semantics across the iterations, which helps the learning of scale-invariant feature representations. Based on ResNet-50, RM largely improves classification accuracy by  $\sim 3.2\%$  on CIFAR-100 and  $\sim 2.8\%$  on ImageNet with negligible extra computation/storage costs. In the downstream object detection task, the RM pretrained model outperforms the baseline by 2.1 AP points and surpasses CutMix by 1.4 AP points under the ATSS detector on COCO. In semantic segmentation, RM also surpasses the baseline and CutMix by 1.9 and 1.1 mIoU points under UperNet on ADE20K, respectively.

## [Bayesian Spline Learning for Equation Discovery of Nonlinear Dynamics with Quantified Uncertainty](#)

- Luning Sun · Daniel Huang · Hao Sun · Jian-Xun Wang
- abstract@[open-review](#): Nonlinear dynamics are ubiquitous in science and engineering applications, but the physics of most complex systems is far from being fully understood. Discovering interpretable governing equations from measurement data can help us understand and predict the behavior of complex dynamic systems. Although extensive work has recently been done in this field, robustly distilling explicit model forms from very sparse data with considerable noise remains intractable. Moreover, quantifying and propagating the uncertainty of the identified system from noisy data is challenging, and relevant literature is still limited. To bridge this gap, we develop a novel Bayesian spline learning framework to identify parsimonious governing equations of nonlinear (spatio)temporal dynamics from sparse, noisy data with quantified uncertainty. The proposed method utilizes spline basis to handle the data scarcity and measurement noise, upon which a group of derivatives can be accurately computed to form a library of candidate model terms. The equation residuals are used to inform the spline learning in a Bayesian manner, where approximate Bayesian uncertainty calibration techniques are employed to approximate posterior distributions of the trainable parameters. To promote the sparsity, an iterative sequential-threshold Bayesian learning approach is developed, using the alternative direction optimization strategy to systematically approximate L0 sparsity constraints. The proposed algorithm is evaluated on multiple nonlinear dynamical systems governed by canonical ordinary and partial differential equations, and the merit/superiority of the proposed method is demonstrated by comparison with state-of-the-art methods.

## [Improving Self-Supervised Learning by Characterizing Idealized Representations](#)

- Yann Dubois · Stefano Ermon · Tatsunori Hashimoto · Percy Liang
- abstract@[open-review](#): Despite the empirical successes of self-supervised learning (SSL) methods, it is unclear what characteristics of their representations lead to high downstream accuracies. In this work, we characterize properties that SSL representations should ideally satisfy. Specifically, we prove necessary and sufficient conditions such that for any task invariant to given data augmentations, probes (e.g., linear or MLP) trained on that representation attain perfect accuracy. These requirements lead to a unifying conceptual framework for improving existing SSL methods and deriving new ones. For contrastive learning, our framework prescribes simple but significant improvements to previous methods such as using asymmetric projection heads. For non-contrastive learning, we use our framework to derive a simple and novel objective. Our resulting SSL algorithms outperform baselines on standard benchmarks, including SwAV+multicrops on linear probing of ImageNet.

## [UniGAN: Reducing Mode Collapse in GANs using a Uniform Generator](#)

- Ziqi Pan · Li Niu · Liqing Zhang
- abstract@[open-review](#): Despite the significant progress that has been made in the training of Generative Adversarial Networks (GANs), the mode collapse problem remains a major challenge in training GANs, which refers to a lack of diversity in generative samples. In this paper, we propose a new type of generative diversity named uniform diversity, which relates to a newly proposed type of mode collapse named  $u$ -mode collapse where the generative samples distribute nonuniformly over the data manifold. From a geometric perspective, we show that the uniform diversity is closely related with the generator uniformity property, and the maximum uniform diversity is achieved if the generator is uniform. To learn a uniform generator, we propose UniGAN, a generative framework with a Normalizing Flow based generator and a simple yet sample efficient generator uniformity regularization, which can be easily adapted to any other generative framework. A new type of diversity metric named udiv is also proposed to estimate the uniform diversity given a set of generative samples in practice. Experimental results verify the effectiveness of our UniGAN in learning a uniform generator and improving uniform diversity.

## [Maximum a posteriori natural scene reconstruction from retinal ganglion cells with deep denoiser priors](#)

- Eric Wu · Alexander Sher · Alan Litke · Eero Simoncelli · E.J. Chichilnisky · Nora Brackbill
- abstract@[open-review](#): A fraction of the visual information arriving at the retina is transmitted to the brain by signals in the optic nerve, and the brain must rely solely on these signals to make inferences about the visual world. Previous work has probed the visual information contained in retinal signals by reconstructing images from retinal activity using linear regression and nonlinear regression with neural networks. Maximum a posteriori (MAP) reconstruction offers a more general and principled approach. We develop a novel method for approximate MAP reconstruction by combining a generalized linear model of light responses in retinal neurons and their dependence on spike history and spikes of neighboring cells, with an image prior implicitly embedded in a deep convolutional neural network trained for image denoising. We use this method to reconstruct natural images from ex vivo simultaneously-recorded spikes of hundreds of ganglion cells uniformly sampling a region of the retina. The method produces reconstructions that match or exceed the state-of-the-art in perceptual similarity and exhibit additional fine detail, while using substantially fewer model parameters than previous approaches. The use of more rudimentary encoding models (a linear-nonlinear-Poisson cascade) or image priors (a 1/F spectral model) significantly reduces reconstruction performance, indicating the essential role of both components in achieving high-quality reconstructed images from the retinal signal.

## [A Theoretical Framework for Inference Learning](#)

- Nick Alonso · Beren Millidge · Jeffrey Krichmar · Emre O Neftci
- abstract@[open-review](#): Backpropagation (BP) is the most successful and widely used algorithm in deep learning. However, the computations required by BP are challenging to reconcile with known neurobiology. This difficulty has stimulated interest in more biologically plausible alternatives to BP. One

such algorithm is the inference learning algorithm (IL). IL has close connections to models of cortical circuits and has achieved equal performance to BP on supervised and auto-associative tasks. In contrast to BP, however, the mathematical foundations of IL are not well-understood. Here, we develop a novel theoretical framework for IL. Our main result is that IL closely approximates an optimization method known as implicit stochastic gradient descent (implicit SGD), which is distinct from the explicit SGD implemented by BP. Our results further show how the standard implementation of IL can be altered to better approximate implicit SGD. Our novel implementation considerably improves the stability of IL across learning rates, which is consistent with our theory, as a key property of implicit SGD is its stability. We provide extensive simulation results that further support our theoretical interpretations and find IL achieves quicker convergence when trained with mini-batch size one while performing competitively with BP for larger mini-batches when combined with Adam.

## [Self-Organized Group for Cooperative Multi-agent Reinforcement Learning](#)

- Jianzhun Shao · Zhiqiang Lou · Hongchang Zhang · Yuhang Jiang · Shuncheng He · Xiangyang Ji
- abstract@[open-review](#): Centralized training with decentralized execution (CTDE) has achieved great success in cooperative multi-agent reinforcement learning (MARL) in practical applications. However, CTDE-based methods typically suffer from poor zero-shot generalization ability with dynamic team composition and varying partial observability. To tackle these issues, we propose a spontaneously grouping mechanism, termed Self-Organized Group (SOG), which is featured with conductor election (CE) and message summary (MS). In CE, a certain number of conductors are elected every  $T$  time-steps to temporally construct groups, each with conductor-follower consensus where the followers are constrained to only communicate with their conductor. In MS, each conductor summarize and distribute the received messages to all affiliate group members to hold a unified scheduling. SOG provides zero-shot generalization ability to the dynamic number of agents and the varying partial observability. Sufficient experiments on mainstream multi-agent benchmarks exhibit superiority of SOG.

## [Retrieve, Reason, and Refine: Generating Accurate and Faithful Patient Instructions](#)

- Fenglin Liu · Bang Yang · Chenyu You · Xian Wu · Shen Ge · Zhangdaihong Liu · Xu Sun · Yang Yang · David Clifton
- abstract@[open-review](#): The "Patient Instruction" (PI), which contains critical instructional information provided both to carers and to the patient at the time of discharge, is essential for the patient to manage their condition outside hospital. An accurate and easy-to-follow PI can improve the self-management of patients which can in turn reduce hospital readmission rates. However, writing an appropriate PI can be extremely time consuming for physicians, and is subject to being incomplete or error-prone for (potentially overworked) physicians. Therefore, we propose a new task that can provide an objective means of avoiding incompleteness, while reducing clinical workload: the automatic generation of the PI, which is imagined as being a document that the clinician can review, modify, and approve as necessary (rather than taking the human "out of the loop"). We build a benchmark clinical dataset and propose the Re\$^3Writer, which imitates the working patterns of physicians to first retrieve related working experience from historical PIs written by physicians, then reason related medical knowledge. Finally, it refines the retrieved working experience and reasoned medical knowledge to extract useful information, which is used to generate the PI for previously-unseen patient according to their health records during hospitalization. Our experiments show that, using our method, the performance of 6 different models can be substantially boosted across all metrics, with up to 20%, 11%, and 19% relative improvements in BLEU-4, ROUGE-L, and METEOR, respectively. Meanwhile, we show results from human evaluations to measure the effectiveness in terms of its usefulness for clinical practice.

## [First Contact: Unsupervised Human-Machine Co-Adaptation via Mutual Information Maximization](#)

- Siddharth Reddy · Sergey Levine · Anca Dragan
- abstract@[open-review](#): How can we train an assistive human-machine interface (e.g., an electromyography-based limb prosthesis) to translate a user's raw command signals into the actions of a robot or computer when there is no prior mapping, we cannot ask the user for supervision in the form of action labels or reward feedback, and we do not have prior knowledge of the tasks the user is trying to accomplish? The key idea in this paper is that, regardless of the task, when an interface is more intuitive, the user's commands are less noisy. We formalize this idea as a completely unsupervised objective for optimizing interfaces: the mutual information between the user's command signals and the induced state transitions in the environment. To evaluate whether this mutual information score can distinguish between effective and ineffective interfaces, we conduct a large-scale observational study on 540K examples of users operating various keyboard and eye gaze interfaces for typing, controlling simulated robots, and playing video games. The results show that our mutual information scores are predictive of the ground-truth task completion metrics in a variety of domains, with an average Spearman's rank correlation of 0.43. In addition to offline evaluation of existing interfaces, we use our unsupervised objective to learn an interface from scratch: we randomly initialize the interface, have the user attempt to perform their desired tasks using the interface, measure the mutual information score, and update the interface to maximize mutual information through reinforcement learning. We evaluate our method through a small-scale user study with 12 participants who perform a 2D cursor control task using a perturbed mouse, and an experiment with one expert user playing the Lunar Lander game using hand gestures captured by a webcam. The results show that we can learn an interface from scratch, without any user supervision or prior knowledge of tasks, with less than 30 minutes of human-in-the-loop training.

## [Interpreting Operation Selection in Differentiable Architecture Search: A Perspective from Influence-Directed Explanations](#)

- Miao Zhang · Wei Huang · Bin Yang
- abstract@[open-review](#): The Differentiable ARchiTecture Search (DARTS) has dominated the neural architecture search community due to its search efficiency and simplicity. DARTS leverages continuous relaxation to convert the intractable operation selection problem into a continuous magnitude optimization problem which can be easily handled with gradient-descent, while it poses an additional challenge in measuring the operation importance or selecting an architecture from the optimized magnitudes. The vanilla DARTS assumes the optimized magnitudes reflect the importance of operations, while more recent works find this naive assumption leads to poor generalization and is without any theoretical guarantees. In this work, we leverage influence functions, the functional derivatives of the loss function, to theoretically reveal the operation selection part in DARTS and estimate the candidate operation importance by approximating its influence on the supernet with Taylor expansions. We show the operation strength is not only related to the magnitude but also second-order information, leading to a fundamentally new criterion for operation selection in DARTS, named Influential Magnitude. Empirical studies across different tasks on several spaces show that vanilla DARTS and its variants can avoid most failures by leveraging the proposed theory-driven operation selection criterion.

## [ALIFE: Adaptive Logit Regularizer and Feature Replay for Incremental Semantic Segmentation](#)

- Youngmin Oh · Donghyeon Baek · Bumsub Ham
- abstract@[open-review](#): We address the problem of incremental semantic segmentation (ISS) recognizing novel object/stuff categories continually without forgetting previous ones that have been learned. The catastrophic forgetting problem is particularly severe in ISS, since pixel-level ground-truth labels are available only for the novel categories at training time. To address the problem, regularization-based methods exploit probability calibration techniques to learn semantic information from unlabeled pixels. While such techniques are effective, there is still lack of theoretical support. Replay-based methods propose to memorize a small set of images for previous categories. They achieve state-of-the-art performance at the cost of large memory footprint. We propose in this paper a novel ISS method, dubbed ALIFE, that provides a better compromise between accuracy and efficiency. To this end, we first show an in-depth analysis on the calibration techniques to better understand the effects on ISS. Based on this, we then introduce an adaptive logit regularizer (ALI) that enables our model to better learn new categories, while retaining knowledge for previous ones. We also present a feature replay scheme that memorizes features, instead of images directly, in order to reduce memory requirements significantly. Since a feature extractor is changed continually, memorized features should also be updated at every incremental stage. To handle this, we introduce category-specific rotation matrices updating the

features for each category separately. We demonstrate the effectiveness of our approach with extensive experiments on standard ISS benchmarks, and show that our method achieves a better trade-off in terms of accuracy and efficiency. Our code and model will be made publicly available online.

## [Probing Classifiers are Unreliable for Concept Removal and Detection](#)

- Abhinav Kumar · Chenhao Tan · Amit Sharma
- abstract@[open-review](#): Neural network models trained on text data have been found to encode undesired linguistic or sensitive attributes in their representation. Removing such attributes is non-trivial because of a complex relationship between the attribute, text input, and the learnt representation. Recent work has proposed post-hoc and adversarial methods to remove such unwanted attributes from a model's representation. Through an extensive theoretical and empirical analysis, we show that these methods can be counter-productive: they are unable to remove the attributes entirely, and in the worst case may end up destroying all task-relevant features. The reason is the methods' reliance on a probing classifier as a proxy for the attribute. Even under the most favorable conditions when an attribute's features in representation space can alone provide 100% accuracy for learning the probing classifier, we prove that post-hoc or adversarial methods will fail to remove the attribute correctly. These theoretical implications are confirmed by empirical experiments on models trained on synthetic, Multi-NLI, and Twitter datasets. For sensitive applications of attribute removal such as fairness, we recommend caution against using these methods and propose a spuriousness metric to gauge the quality of the final classifier.

## [Confident Adaptive Language Modeling](#)

- Tal Schuster · Adam Fisch · Jai Gupta · Mostafa Dehghani · Dara Bahri · Vinh Tran · Yi Tay · Donald Metzler
- abstract@[open-review](#): Recent advances in Transformer-based large language models (LLMs) have led to significant performance improvements across many tasks. These gains come with a drastic increase in the models' size, potentially leading to slow and costly use at inference time. In practice, however, the series of generations made by LLMs is composed of varying levels of difficulty. While certain predictions truly benefit from the models' full capacity, other continuations are more trivial and can be solved with reduced compute. In this work, we introduce Confident Adaptive Language Modeling (CALM), a framework for dynamically allocating different amounts of compute per input and generation timestep. Early exit decoding involves several challenges that we address here, such as: (1) what confidence measure to use; (2) connecting sequence-level constraints to local per-token exit decisions; and (3) attending back to missing hidden representations due to early exits in previous tokens. Through theoretical analysis and empirical experiments on three diverse text generation tasks, we demonstrate the efficacy of our framework in reducing compute---potential speedup of up to  $\times 3$ ---while provably maintaining high performance.

## [Weighted Mutual Learning with Diversity-Driven Model Compression](#)

- Miao Zhang · Li Wang · David Campos · Wei Huang · Chenjuan Guo · Bin Yang
- abstract@[open-review](#): Online distillation attracts attention from the community as it simplifies the traditional two-stage knowledge distillation process into a single stage. Online distillation collaboratively trains a group of peer models, which are treated as students, and all students gain extra knowledge from each other. However, memory consumption and diversity among peers are two key challenges to the scalability and quality of online distillation. To address the two challenges, this paper presents a framework called Weighted Mutual Learning with Diversity-Driven Model Compression (\textbf{WML}) for online distillation. First, at the base of a hierarchical structure where peers share different parts, we leverage the structured network pruning to generate diversified peer models and reduce the memory requirements. Second, rather than taking the average of peers, this paper, for the first time, leverages a bi-level formulation to estimate the relative importance of peers with a close-form, to further boost the effectiveness of the distillation from each other. Extensive experiments show the generalization of the proposed framework, which outperforms existing online distillation methods on a variety of deep neural networks. More interesting, as a byproduct, \textbf{WML} produces a series of pruned models under different model sizes in a single run, which also achieves competitive results compared with existing channel pruning methods.

## [Beyond the Best: Distribution Functional Estimation in Infinite-Armed Bandits](#)

- Yifei Wang · Tavor Baharav · Yanjun Han · Jiantao Jiao · David Tse
- abstract@[open-review](#): In the infinite-armed bandit problem, each arm's average reward is sampled from an unknown distribution, and each arm can be sampled further to obtain noisy estimates of the average reward of that arm. Prior work focuses on the best arm, i.e. estimating the maximum of the average reward distribution. We consider a general class of distribution functionals beyond the maximum and obtain optimal sample complexities in both offline and online settings. We show that online estimation, where the learner can sequentially choose whether to sample a new or existing arm, offers no advantage over the offline setting for estimating the mean functional, but significantly reduces the sample complexity for other functionals such as the median, maximum, and trimmed mean. We propose unified meta algorithms for the online and offline settings and derive matching lower bounds using different Wasserstein distances. For the special case of median estimation, we identify a curious thresholding phenomenon on the indistinguishability between Gaussian convolutions with respect to the noise level, which may be of independent interest.

## [On the inability of Gaussian process regression to optimally learn compositional functions](#)

- Matteo Giordano · Kolyan Ray · Johannes Schmidt-Hieber
- abstract@[open-review](#): We rigorously prove that deep Gaussian process priors can outperform Gaussian process priors if the target function has a compositional structure. To this end, we study information-theoretic lower bounds for posterior contraction rates for Gaussian process regression in a continuous regression model. We show that if the true function is a generalized additive function, then the posterior based on any mean-zero Gaussian process can only recover the truth at a rate that is strictly slower than the minimax rate by a factor that is polynomially suboptimal in the sample size  $n$ .

## [Pessimism for Offline Linear Contextual Bandits using \$\ell\_p\$ Confidence Sets](#)

- Gene Li · Cong Ma · Nati Srebro
- abstract@[open-review](#): We present a family  $\{\widehat{\pi}_p\}_{p \geq 1}$  of pessimistic learning rules for offline learning of linear contextual bandits, relying on confidence sets with respect to different  $\ell_p$  norms, where  $\widehat{\pi}_2$  corresponds to Bellman-consistent pessimism (BCP), while  $\widehat{\pi}_{\infty}$  is a novel generalization of lower confidence bound (LCB) to the linear setting. We show that the novel  $\widehat{\pi}_{\infty}$  learning rule is, in a sense, adaptively optimal, as it achieves the minimax performance (up to log factors) against all  $\ell_q$ -constrained problems, and as such it strictly dominates all other predictors in the family, including  $\widehat{\pi}_2$ .

## [One Inlier is Enough: Towards Efficient Position Encoding for Point Cloud Registration](#)

- Fan Yang · Lin Guo · Zhi Chen · Wenbing Tao
- abstract@[open-review](#): Transformer architecture has shown great potential for many visual tasks, including point cloud registration. As an order-aware module, position encoding plays an important role in Transformer architecture applied to point cloud registration task. In this paper, we propose a one-inlier based position encoding method for point cloud registration network. Specifically, we first find one correspondence by a differentiable optimal transport layer, and use it to normalize each point for position encoding. It can eliminate the challenges brought by the different reference frames of two point clouds, and mitigate the feature ambiguity by learning the spatial consistency. Then, we propose a joint approach for establishing correspondence and position encoding, presenting an iterative optimization process. Finally, we design a progressive way for point cloud alignment and feature learning to

gradually optimize the rigid transformation. The proposed position encoding is very efficient, requiring only a small addition of memory and computing overhead. Extensive experiments demonstrate the proposed method can achieve competitive performance with the state-of-the-art methods in both indoor and outdoor scenes.

## [Non-Markovian Reward Modelling from Trajectory Labels via Interpretable Multiple Instance Learning](#)

- Joseph Early · Tom Bewley · Christine Evers · Sarvapali Ramchurn
- abstract@[open-review](#): We generalise the problem of reward modelling (RM) for reinforcement learning (RL) to handle non-Markovian rewards. Existing work assumes that human evaluators observe each step in a trajectory independently when providing feedback on agent behaviour. In this work, we remove this assumption, extending RM to capture temporal dependencies in human assessment of trajectories. We show how RM can be approached as a multiple instance learning (MIL) problem, where trajectories are treated as bags with return labels, and steps within the trajectories are instances with unseen reward labels. We go on to develop new MIL models that are able to capture the time dependencies in labelled trajectories. We demonstrate on a range of RL tasks that our novel MIL models can reconstruct reward functions to a high level of accuracy, and can be used to train high-performing agent policies.

## [Faster Stochastic Algorithms for Minimax Optimization under Polyak-{L}ojasiewicz Condition](#)

- Lesi Chen · Boyuan Yao · Luo Luo
- abstract@[open-review](#): This paper considers stochastic first-order algorithms for minimax optimization under Polyak-{L}ojasiewicz (PL) conditions. We propose SPIDER-GDA for solving the finite-sum problem of the form  $\min_x \max_y f(x,y)$ , where the objective function  $f(x,y)$  is  $\mu_x$ -PL in  $x$  and  $\mu_y$ -PL in  $y$ ; and each  $f_i(x,y)$  is  $L$ -smooth. We prove SPIDER-GDA could find an  $\epsilon$ -approximate solution within  $O((n + \sqrt{n})\kappa_x\kappa_y^2\log(1/\epsilon))$  stochastic first-order oracle (SFO) complexity, which is better than the state-of-the-art method whose SFO upper bound is  $O((n + n^{2/3})\kappa_x\kappa_y^2\log(1/\epsilon))$ , where  $\kappa_x \triangleq L/\mu_x$  and  $\kappa_y \triangleq L/\mu_y$ . For the ill-conditioned case, we provide an accelerated algorithm to reduce the computational cost further. It achieves  $\tilde{O}((n + \sqrt{n})\kappa_x\kappa_y)\log^2(1/\epsilon)$  SFO upper bound when  $\kappa_x \geq \sqrt{n}$ . Our ideas also can be applied to the more general setting that the objective function only satisfies PL condition for one variable. Numerical experiments validate the superiority of proposed methods.

## [Class-Aware Generative Adversarial Transformers for Medical Image Segmentation](#)

- Chenyu You · Ruihan Zhao · Fenglin Liu · Siyuan Dong · Sandeep Chinchali · Ufuk Topcu · Lawrence Staib · James Duncan
- abstract@[open-review](#): Transformers have made remarkable progress towards modeling long-range dependencies within the medical image analysis domain. However, current transformer-based models suffer from several disadvantages: (1) existing methods fail to capture the important features of the images due to the naive tokenization scheme; (2) the models suffer from information loss because they only consider single-scale feature representations; and (3) the segmentation label maps generated by the models are not accurate enough without considering rich semantic contexts and anatomical textures. In this work, we present CASTformer, a novel type of adversarial transformers, for 2D medical image segmentation. First, we take advantage of the pyramid structure to construct multi-scale representations and handle multi-scale variations. We then design a novel class-aware transformer module to better learn the discriminative regions of objects with semantic structures. Lastly, we utilize an adversarial training strategy that boosts segmentation accuracy and correspondingly allows a transformer-based discriminator to capture high-level semantically correlated contents and low-level anatomical features. Our experiments demonstrate that CASTformer dramatically outperforms previous state-of-the-art transformer-based approaches on three benchmarks, obtaining 2.54%-5.88% absolute improvements in Dice over previous models. Further qualitative experiments provide a more detailed picture of the model's inner workings, shed light on the challenges in improved transparency, and demonstrate that transfer learning can greatly improve performance and reduce the size of medical image datasets in training, making CASTformer a strong starting point for downstream medical image analysis tasks. Codes and models will be made available to public.

## [Divert More Attention to Vision-Language Tracking](#)

- Mingzhe Guo · Zhipeng Zhang · Heng Fan · Liping Jing
- abstract@[open-review](#): Relying on Transformer for complex visual feature learning, object tracking has witnessed the new standard for state-of-the-arts (SOTAs). However, this advancement accompanies by larger training data and longer training period, making tracking increasingly expensive. In this paper, we demonstrate that the Transformer-reliance is not necessary and the pure ConvNets are still competitive and even better yet more economical and friendly in achieving SOTA tracking. Our solution is to unleash the power of multimodal vision-language (VL) tracking, simply using ConvNets. The essence lies in learning novel unified-adaptive VL representations with our modality mixer (ModaMixer) and asymmetrical ConvNet search. We show that our unified-adaptive VL representation, learned purely with the ConvNets, is a simple yet strong alternative to Transformer visual features, by unbelievably improving a CNN-based Siamese tracker by 14.5% in SUC on challenging LaSOT (50.7% → 65.2%), even outperforming several Transformer-based SOTA trackers. Besides empirical results, we theoretically analyze our approach to evidence its effectiveness. By revealing the potential of VL representation, we expect the community to divert more attention to VL tracking and hope to open more possibilities for future tracking beyond Transformer. Code and models are released at <https://github.com/JudasDie/SOTS>.

## [Towards Efficient 3D Object Detection with Knowledge Distillation](#)

- Jihan Yang · Shaoshuai Shi · Runyu Ding · Zhe Wang · Xiaojuan Qi
- abstract@[open-review](#): Despite substantial progress in 3D object detection, advanced 3D detectors often suffer from heavy computation overheads. To this end, we explore the potential of knowledge distillation (KD) for developing efficient 3D object detectors, focusing on popular pillar- and voxel-based detectors. Without well-developed teacher-student pairs, we first study how to obtain student models with good trade offs between accuracy and efficiency from the perspectives of model compression and input resolution reduction. Then, we build a benchmark to assess existing KD methods developed in the 2D domain for 3D object detection upon six well-constructed teacher-student pairs. Further, we propose an improved KD pipeline incorporating an enhanced logit KD method that performs KD on only a few pivotal positions determined by teacher classification response and a teacher-guided student model initialization to facilitate transferring teacher model's feature extraction ability to students through weight inheritance. Finally, we conduct extensive experiments on the Waymo dataset. Our best performing model achieves 65.75% LEVEL 2 mAPH surpassing its teacher model and requiring only 44% of teacher flops. Our most efficient model runs 51 FPS on an NVIDIA A100, which is 2.2times faster than PointPillar with even higher accuracy. Code will be available.

## [GAMA: Generative Adversarial Multi-Object Scene Attacks](#)

- Abhishek Aich · Calvin-Khang Ta · Akash Gupta · Chengyu Song · Srikanth Krishnamurthy · Salman Asif · Amit Roy-Chowdhury
- abstract@[open-review](#): The majority of methods for crafting adversarial attacks have focused on scenes with a single dominant object (e.g., images from ImageNet). On the other hand, natural scenes include multiple dominant objects that are semantically related. Thus, it is crucial to explore designing attack strategies that look beyond learning on single-object scenes or attack single-object victim classifiers. Due to their inherent property of strong transferability of perturbations to unknown models, this paper presents the first approach of using generative models for adversarial attacks on multi-object scenes. In order to represent the relationships between different objects in the input scene, we leverage upon the open-sourced pre-trained vision-language model CLIP (Contrastive Language-Image Pre-training), with the motivation to exploit the encoded semantics in the language space along with

the visual space. We call this attack approach Generative Adversarial Multi-object scene Attacks (GAMA). For the first time in literature, GAMA demonstrates the utility of the CLIP model as an attacker's tool in order to train formidable perturbation generators. Using the joint image-text features to train the generator, we show that GAMA can craft potent transferable perturbations in order to fool victim classifiers in various attack settings. For example, GAMA triggers  $\sim 16\%$  more misclassification than state-of-the-art generative approaches in black-box settings where both the classifier architecture and data distribution of the attacker are different from the victim. Our code will be made publicly available.

## [Adversarially Robust Learning: A Generic Minimax Optimal Learner and Characterization](#)

- Omar Montasser · Steve Hanneke · Nati Srebro
- abstract@[open-review](#): We present a minimax optimal learner for the problem of learning predictors robust to adversarial examples at test-time. Interestingly, we find that this requires new algorithmic ideas and approaches to adversarially robust learning. In particular, we show, in a strong negative sense, the suboptimality of the robust learner proposed by Montasser, Hanneke, and Srebro [2019] and a broader family of learners we identify as local learners. Our results are enabled by adopting a global perspective, specifically, through a key technical contribution: the global one-inclusion graph, which may be of independent interest, that generalizes the classical one-inclusion graph due to Haussler, Littlestone, and Warmuth [1994]. Finally, as a byproduct, we identify a dimension characterizing qualitatively and quantitatively what classes of predictors  $\mathcal{H}$  are robustly learnable. This resolves an open problem due to Montasser et al. [2019], and closes a (potentially) infinite gap between the established upper and lower bounds on the sample complexity of adversarially robust learning.

## [End-to-end Stochastic Programming with Energy-based Model](#)

- Lingkai Kong · Jiaming Cui · Yuchen Zhuang · Rui Feng · B. Aditya Prakash · Chao Zhang
- abstract@[open-review](#): Solving optimization problems with unknown parameters often requires learning a predictive model to predict the distribution of the unknown parameters and then solving the stochastic problem using these values. However, the criteria by which the predictive model is trained are often inconsistent with the goal of the downstream optimization task. Decision focused learning has been proposed to directly incorporate the optimization objective into training. However, it has poor scalability since it requires to solve and differentiate through the optimization problem at every iteration; furthermore, it can only be applied to convex problem. To address these shortcomings, we propose a new end-to-end stochastic programming method with Energy-based model. The core of our method is to directly model the probability of the decision variable conditioned on the input features using the energy parameterization. To leverage the algorithmic structure of the optimization problem, we parameterize the energy function with the expected downstream task loss; To capture the both the optimum location and overall energy shape, we augment the maximum likelihood training objective with a distribution based regularizer. We evaluate our method in three applications: energy scheduling, covid-19 resource allocation and non-convex adverarial security game, demonstrating that our method outperforms existing methods with a large reduction in training time.

## [Memorization and Optimization in Deep Neural Networks with Minimum Over-parameterization](#)

- Simone Bombari · Mohammad Hossein Amani · Marco Mondelli
- abstract@[open-review](#): The Neural Tangent Kernel (NTK) has emerged as a powerful tool to provide memorization, optimization and generalization guarantees in deep neural networks. A line of work has studied the NTK spectrum for two-layer and deep networks with at least a layer with  $\Omega(N)$  neurons,  $N$  being the number of training samples. Furthermore, there is increasing evidence suggesting that deep networks with sub-linear layer widths are powerful memorizers and optimizers, as long as the number of parameters exceeds the number of samples. Thus, a natural open question is whether the NTK is well conditioned in such a challenging sub-linear setup. In this paper, we answer this question in the affirmative. Our key technical contribution is a lower bound on the smallest NTK eigenvalue for deep networks with the minimum possible over-parameterization: up to logarithmic factors, the number of parameters is  $\Omega(N)$  and, hence, the number of neurons is as little as  $\sqrt{N}$ . To showcase the applicability of our NTK bounds, we provide two results concerning memorization capacity and optimization guarantees for gradient descent training.

## [UnfoldML: A Cost-Aware 2-D Dynamic Prediction Pipeline for Multi-Stage Classification](#)

- Yanbo Xu · Alind Khare · Glenn Matlin · Monish Ramadoss · Rishikesan Kamaleswaran · Chao Zhang · Alexey Tumanov
- abstract@[open-review](#): Prior focus on Machine learning (ML) models has been on maximizing the accuracy of predictive tasks. A new trend emerges, however---ML models become increasingly more complex, resource intensive, and costlier to deploy in resource constrained run time environments. This issue is exacerbated for prediction tasks involving sequential classification on progressively transitioned stages, characterized as happens-before" relationship between stages, with increasing data volume and data modalities. We make an observation that it is possible to "unfold" a monolithic single multi-class classifier, typically trained for all stages using all data, into a series of single-stage classifiers. Then, each single-stage classifier can be cascaded gradually from cheaper to more expensive binary classifiers that are trained using only the necessary data modalities or features required for that stage. This observation led us to develop UnfoldML---a cost-aware 2-D dynamic prediction pipeline for multi-stage classification. Through the combination of this 2-D query propagation mechanism and a set of policies for dynamic model selection, UnfoldML is able to (1) navigate an accuracy/cost tradeoff space, (2) reduce the spatio-temporal cost of inference by orders of magnitude, and (3) enable early prediction on succeeding stages. UnfoldML can benefit a wide range of real-world problems. We demonstrate UnfoldML in clinical settings where it can detect multi-stage disease development in real time. It achieves within 0.1% accuracy from the highest-performing multi-class baseline, while saving close to 20X on spatio-temporal cost of inference and detecting about 3.5 hour earlier on the second stage. We also demonstrate UnfoldML in image classifications where it can predict different level of labels (from coarse to fine) given the different level of abstractions of a image. It achieves close to a 5X cost reduction with as little as 0.4% accuracy reduction in fine label prediction. Overall, UnfoldML improves the whole frontier of optimality in the accuracy/cost tradeoff space, achieving the highest area under the curve.

## [A Unified Model for Multi-class Anomaly Detection](#)

- Zhiyuan You · Lei Cui · Yujun Shen · Kai Yang · Xin Lu · Yu Zheng · Xinyi Le
- abstract@[open-review](#): Despite the rapid advance of unsupervised anomaly detection, existing methods require to train separate models for different objects. In this work, we present UniAD that accomplishes anomaly detection for multiple classes with a unified framework. Under such a challenging setting, popular reconstruction networks may fall into an "identical shortcut", where both normal and anomalous samples can be well recovered, and hence fail to spot outliers. To tackle this obstacle, we make three improvements. First, we revisit the formulations of fully-connected layer, convolutional layer, as well as attention layer, and confirm the important role of query embedding (i.e., within attention layer) in preventing the network from learning the shortcut. We therefore come up with a layer-wise query decoder to help model the multi-class distribution. Second, we employ a neighbor masked attention module to further avoid the information leak from the input feature to the reconstructed output feature. Third, we propose a feature jittering strategy that urges the model to recover the correct message even with noisy inputs. We evaluate our algorithm on MVTec-AD and CIFAR-10 datasets, where we surpass the state-of-the-art alternatives by a sufficiently large margin. For example, when learning a unified model for 15 categories in MVTec-AD, we surpass the second competitor on the tasks of both anomaly detection (from 88.1% to 96.5%) and anomaly localization (from 89.5% to 96.8%). Code will be made publicly available.

## [Online Allocation and Learning in the Presence of Strategic Agents](#)

- Steven Yin · Shipra Agrawal · Assaf Zeevi

- abstract@[open-review](#): We study the problem of allocating \$T\$ sequentially arriving items among \$n\$ homogenous agents under the constraint that each agent must receive a prespecified fraction of all items, with the objective of maximizing the agents' total valuation of items allocated to them. The agents' valuations for the item in each round are assumed to be i.i.d. but their distribution is apriori unknown to the central planner. Therefore, the central planner needs to implicitly learn these distributions from the observed values in order to pick a good allocation policy. However, an added challenge here is that the agents are strategic with incentives to misreport their valuations in order to receive better allocations. This sets our work apart both from the online auction mechanism design settings which typically assume known valuation distributions and/or involve payments, and from the online learning settings that do not consider strategic agents. To that end, our main contribution is an online learning based allocation mechanism that is approximately Bayesian incentive compatible, and when all agents are truthful, guarantees a sublinear regret for individual agents' utility compared to that under the optimal offline allocation policy.

## [On the role of overparameterization in Temporal Difference learning with linear function approximation](#)

- Valentin Thomas
- abstract@[open-review](#): Much of the recent successes of deep learning can be attributed to scaling up the size of the networks to the point where they often are vastly overparameterized. Thus, understanding the role of overparameterization is of increasing importance. While predictive theories have been developed for supervised learning, little is known about the Reinforcement Learning case. In this work, we take a theoretical approach and study the role of overparameterization for off-policy Temporal Difference (TD) learning in the linear setting. We leverage tools from Random Matrix Theory and random graph theory to obtain a characterization of the spectrum of the TD operator. We use this result to study the stability and optimization dynamics of TD learning as a function of the number of parameters.

## [What are the best Systems? New Perspectives on NLP Benchmarking](#)

- Pierre Colombo · Nathan Noiry · Ekhine Irurozki · Stephan Clomençon
- abstract@[open-review](#): In Machine Learning, a benchmark refers to an ensemble of datasets associated with one or multiple metrics together with a way to aggregate different systems performances. They are instrumental in {it (i)} assessing the progress of new methods along different axes and {it (ii)} selecting the best systems for practical use. This is particularly the case for NLP with the development of large pre-trained models (\textit{e.g.} GPT, BERT) that are expected to generalize well on a variety of tasks. While the community mainly focused on developing new datasets and metrics, there has been little interest in the aggregation procedure, which is often reduced to a simple average over various performance measures. However, this procedure can be problematic when the metrics are on a different scale, which may lead to spurious conclusions. This paper proposes a new procedure to rank systems based on their performance across different tasks. Motivated by the social choice theory, the final system ordering is obtained through aggregating the rankings induced by each task and is theoretically grounded. We conduct extensive numerical experiments (on over 270k scores) to assess the soundness of our approach both on synthetic and real scores (\textit{e.g.} GLUE, EXTREM, SEVAL, TAC, FLICKR). In particular, we show that our method yields different conclusions on state-of-the-art systems than the mean-aggregation procedure while being both more reliable and robust.

## [Scale-invariant Learning by Physics Inversion](#)

- Philipp Holl · Vladlen Koltun · Nils Thuerey
- abstract@[open-review](#): Solving inverse problems, such as parameter estimation and optimal control, is a vital part of science. Many experiments repeatedly collect data and rely on machine learning algorithms to quickly infer solutions to the associated inverse problems. We find that state-of-the-art training techniques are not well-suited to many problems that involve physical processes. The highly nonlinear behavior, common in physical processes, results in strongly varying gradients that lead first-order optimizers like SGD or Adam to compute suboptimal optimization directions. We propose a novel hybrid training approach that combines higher-order optimization methods with machine learning techniques. We take updates from a scale-invariant inverse problem solver and embed them into the gradient-descent-based learning pipeline, replacing the regular gradient of the physical process. We demonstrate the capabilities of our method on a variety of canonical physical systems, showing that it yields significant improvements on a wide range of optimization and learning problems.

## [Learning Interface Conditions in Domain Decomposition Solvers](#)

- Ali Taghibakhshi · Nicolas Nytko · Tareq Uz Zaman · Scott MacLachlan · Luke Olson · Matthew West
- abstract@[open-review](#): Domain decomposition methods are widely used and effective in the approximation of solutions to partial differential equations. Yet the \textit{optimal} construction of these methods requires tedious analysis and is often available only in simplified, structured-grid settings, limiting their use for more complex problems. In this work, we generalize optimized Schwarz domain decomposition methods to unstructured-grid problems, using Graph Convolutional Neural Networks (GCNNs) and unsupervised learning to learn optimal modifications at subdomain interfaces. A key ingredient in our approach is an improved loss function, enabling effective training on relatively small problems, but robust performance on arbitrarily large problems, with computational cost linear in problem size. The performance of the learned linear solvers is compared with both classical and optimized domain decomposition algorithms, for both structured- and unstructured-grid problems.

## [The computational and learning benefits of Daleian neural networks](#)

- Adam Haber · Elad Schneidman
- abstract@[open-review](#): Dale's principle implies that biological neural networks are composed of neurons that are either excitatory or inhibitory. While the number of possible architectures of such Daleian networks is exponentially smaller than the number of non-Daleian ones, the computational and functional implications of using Daleian networks by the brain are mostly unknown. Here, we use models of recurrent spiking neural networks and rate-based ones to show, surprisingly, that despite the structural limitations on Daleian networks, they can approximate the computation performed by non-Daleian networks to a very high degree of accuracy. Moreover, we find that Daleian networks are more functionally robust to synaptic noise. We then show that unlike non-Daleian networks, Daleian ones can learn efficiently by tuning of single neuron features, nearly as well as learning by tuning individual synaptic weights. Importantly, this suggests a simpler and more biologically plausible learning mechanisms. We therefore suggest that in addition to architectural simplicity, Dale's principle confers computational and learning benefits for biological networks, and offer new directions for constructing and training biologically-inspired artificial neural networks.

## [Doubly-Asynchronous Value Iteration: Making Value Iteration Asynchronous in Actions](#)

- Tian Tian · Kenny Young · Richard Sutton
- abstract@[open-review](#): Value iteration (VI) is a foundational dynamic programming method, important for learning and planning in optimal control and reinforcement learning. VI proceeds in batches, where the update to the value of each state must be completed before the next batch of updates can begin. Completing a single batch is prohibitively expensive if the state space is large, rendering VI impractical for many applications. Asynchronous VI helps to address the large state space problem by updating one state at a time, in-place and in an arbitrary order. However, Asynchronous VI still requires a maximization over the entire action space, making it impractical for domains with large action space. To address this issue, we propose \textit{doubly-asynchronous value iteration} (DAVI), a new algorithm that generalizes the idea of asynchrony from states to states and actions. More concretely, DAVI only considers a subset of actions that can be of any user-defined size. In this paper, we establish that DAVI maintains similarly appealing theoretical properties to VI without the need to wait for a full sweep through the entire action space in each update. Specifically, DAVI converges to the optimal value

function with probability one, converges at a near-geometric rate with probability  $1-\delta$ , and returns a near-optimal policy in computation time that nearly matches a previously established bound for VI. We also empirically demonstrate DAVI's effectiveness in several experiments.

## [ShuffleMixer: An Efficient ConvNet for Image Super-Resolution](#)

- Long Sun · Jinshan Pan · Jinhui Tang
- abstract@[open-review](#): Lightweight and efficiency are critical drivers for the practical application of image super-resolution (SR) algorithms. We propose a simple and effective approach, ShuffleMixer, for lightweight image super-resolution that combines large convolution and channel split-shuffle operation. In contrast to previous SR models that stack multiple small kernel convolutions or complex operators to learn representations, we explore a large kernel ConvNet for mobile-friendly SR design. Specifically, we develop a large depthwise convolution and two projection layers based on channel splitting and shuffling as the basic component to mix features efficiently. Since the contexts of natural images are strongly locally correlated, using large depthwise convolutions only is insufficient to reconstruct fine details. To overcome this problem while keeping it lightweight and efficient, we introduce the Fused-MBConv into the proposed network to model the local connectivity of different features. Experimental results demonstrate that the proposed ShuffleMixer is about  $6 \times$  smaller than the state-of-the-art methods in terms of model parameters and FLOPs while achieving competitive performance.

## [Beyond neural scaling laws: beating power law scaling via data pruning](#)

- Ben Sorscher · Robert Geirhos · Shashank Shekhar · Surya Ganguli · Ari Morcos
- abstract@[open-review](#): Widely observed neural scaling laws, in which error falls off as a power of the training set size, model size, or both, have driven substantial performance improvements in deep learning. However, these improvements through scaling alone require considerable costs in compute and energy. Here we focus on the scaling of error with dataset size and show both in theory and practice that we can break beyond power law scaling and reduce it to exponential scaling instead if we have access to a high-quality data pruning metric that ranks the order in which training examples should be discarded to achieve any pruned dataset size. We then test this new exponential scaling prediction with pruned dataset size empirically, and indeed observe better than power-law scaling performance on ResNets trained on CIFAR-10, SVHN, and ImageNet. Given the importance of finding high-quality pruning metrics, we perform the first large-scale benchmarking study of 9 different data pruning metrics on ImageNet. We find most existing high performing metrics scale poorly to ImageNet, while the best are computationally intensive and require labels for every image. We therefore developed a new simple, cheap and scalable self-supervised pruning metric that demonstrates comparable performance to the best supervised metrics. Overall, our work suggests that the discovery of good data-pruning metrics may provide a viable path forward to substantially improved neural scaling laws, thereby reducing the resource costs of modern deep learning.

## [Learning to Sample and Aggregate: Few-shot Reasoning over Temporal Knowledge Graph](#)

- Ruijie Wang · Zheng Li · Dachun Sun · Shengzhong Liu · Jinning Li · Bing Yin · Tarek Abdelzaher
- abstract@[open-review](#): In this paper, we investigate a realistic but underexplored problem, called few-shot temporal knowledge graph reasoning, that aims to predict future facts for newly emerging entities based on extremely limited observations in evolving graphs. It offers practical value in applications that need to derive instant new knowledge about new entities in temporal knowledge graphs (TKGs) with minimal supervision. The challenges mainly come from the few-shot and time shift properties of new entities. First, the limited observations associated with them are insufficient for training a model from scratch. Second, the potentially dynamic distributions from the initially observable facts to the future facts ask for explicitly modeling the evolving characteristics of new entities. We correspondingly propose a novel Meta Temporal Knowledge Graph Reasoning (MetaTKGR) framework. Unlike prior work that relies on rigid neighborhood aggregation schemes to enhance low-data entity representation, MetaTKGR dynamically adjusts the strategies of sampling and aggregating neighbors from recent facts for new entities, through temporally supervised signals on future facts as instant feedback. Besides, such a meta temporal reasoning procedure goes beyond existing meta-learning paradigms on static knowledge graphs that fail to handle temporal adaptation with large entity variance. We further provide a theoretical analysis and propose a temporal adaptation regularizer to stabilize the meta temporal reasoning over time. Empirically, extensive experiments on three real-world TKGs demonstrate the superiority of MetaTKGR over eight state-of-the-art baselines by a large margin.

## [On-Demand Sampling: Learning Optimally from Multiple Distributions](#)

- Nika Haghtalab · Michael Jordan · Eric Zhao
- abstract@[open-review](#): Social and real-world considerations such as robustness, fairness, social welfare, and multi-agent trade-offs have given rise to multi-distribution learning paradigms, such as collaborative [Blum et al. 2017], group distributionally robust [Sagawa et al. 2019], and fair federated [Mohri et al. 2019] learning. In each of these settings, a learner seeks to minimize its worst-case loss over a set of  $n$  predefined distributions, while using as few samples as possible. In this paper, we establish the optimal sample complexity of these learning paradigms and give algorithms that meet this sample complexity. Importantly, our sample complexity bounds exceed that of the sample complexity of learning a single distribution only by an additive factor of  $\frac{n}{\log(n)}\epsilon^2$ . These improve upon the best known sample complexity of agnostic federated learning by Mohri et al. 2019 by a multiplicative factor of  $n$ , the sample complexity of collaborative learning by Nguyen and Zakythinou 2018 by a multiplicative factor of  $\frac{\log(n)}{\log(n)}\epsilon^3$ , and give the first sample complexity bounds for the group DRO objective of Sagawa et al. 2019. To achieve optimal sample complexity, our algorithms learn to sample and learn from distributions on demand. Our algorithm design and analysis is enabled by our extensions of stochastic optimization techniques for solving stochastic zero-sum games. In particular, we contribute variants of Stochastic Mirror Descent that can trade off between players' access to cheap one-off samples and more expensive reusable ones.

## [A permutation-free kernel two-sample test](#)

- Shubhangsu Shekhar · Ilmun Kim · Aaditya Ramdas
- abstract@[open-review](#): The kernel Maximum Mean Discrepancy~(MMD) is a popular multivariate distance metric between distributions. The usual kernel-MMD test statistic (for two-sample testing) is a degenerate U-statistic under the null, and thus it has an intractable limiting null distribution. Hence, the standard approach for designing a level- $(1-\alpha)$  two-sample test using this statistic involves selecting the rejection threshold as the  $(1-\alpha)$ -quantile of the permutation distribution. The resulting nonparametric test has finite-sample validity but suffers from large computational cost, since the test statistic must be recomputed for every permutation. We propose the cross-MMD, a new quadratic time MMD test statistic based on sample-splitting and studentization. We prove that under mild assumptions, it has a standard normal limiting distribution under the null. Importantly, we also show that the resulting test is consistent against any fixed alternative, and has minimax rate-optimal power against local alternatives (with Gaussian kernels). For large sample-sizes, our new cross-MMD provides a significant speedup over the MMD, for only a slight loss in power.

## [Factored Adaptation for Non-Stationary Reinforcement Learning](#)

- Fan Feng · Biwei Huang · Kun Zhang · Sara Magliacane
- abstract@[open-review](#): Dealing with non-stationarity in environments (i.e., transition dynamics) and objectives (i.e., reward functions) is a challenging problem that is crucial in real-world applications of reinforcement learning (RL). While most current approaches model the changes as a single shared embedding vector, we leverage insights from the recent causality literature to model non-stationarity in terms of individual latent change factors and causal graphs across different environments. In particular, we propose Factored Adaptation for Non-Stationary RL (FANS-RL), a factored adaption approach that learns jointly the causal structure in terms of a factored MDP, and a factored representation of the individual time-varying change factors. We prove that under standard assumptions we can recover completely the causal graph representing the factored transition and reward function, and a

partial structure between the individual change factors and the state components. Through our general framework, we can consider general non-stationary scenarios with different changing function types and changing frequency, including changes across episodes and within episodes. Experimental results demonstrate that FANS-RL outperforms existing approaches in terms of rewards, compactness of the latent state representation and robustness to varying degrees of non-stationarity.

## [Feature Learning in \$L\_2\$ -regularized DNNs: Attraction/Repulsion and Sparsity](#)

- Arthur Jacot · Eugene Golikov · Clement Hongler · Franck Gabriel
- abstract@[open-review](#): We study the loss surface of DNNs with  $L_2$  regularization. We show that the loss in terms of the parameters can be reformulated into a loss in terms of the layerwise activations  $Z_{\{\ell\}}$  of the training set. This reformulation reveals the dynamics behind feature learning: each hidden representations  $Z_{\{\ell\}}$  are optimal w.r.t. to an attraction/repulsion problem and interpolate between the input and output representations, keeping as little information from the input as necessary to construct the activation of the next layer. For positively homogeneous non-linearities, the loss can be further reformulated in terms of the covariances of the hidden representations, which takes the form of a partially convex optimization over a convex cone. This second reformulation allows us to prove a sparsity result for homogeneous DNNs: any local minimum of the  $L_2$ -regularized loss can be achieved with at most  $N(N+1)$  neurons in each hidden layer (where  $N$  is the size of the training set). We show that this bound is tight by giving an example of a local minimum that requires  $N^2/4$  hidden neurons. But we also observe numerically that in more traditional settings much less than  $N^2$  neurons are required to reach the minima.

## [Are All Losses Created Equal: A Neural Collapse Perspective](#)

- Jinxin Zhou · Chong You · Xiao Li · Kangning Liu · Sheng Liu · Qing Qu · Zhihui Zhu
- abstract@[open-review](#): While cross entropy (CE) is the most commonly used loss function to train deep neural networks for classification tasks, many alternative losses have been developed to obtain better empirical performance. Among them, which one is the best to use is still a mystery, because there seem to be multiple factors affecting the answer, such as properties of the dataset, the choice of network architecture, and so on. This paper studies the choice of loss function by examining the last-layer features of deep networks, drawing inspiration from a recent line work showing that the global optimal solution of CE and mean-square-error (MSE) losses exhibits a Neural Collapse phenomenon. That is, for sufficiently large networks trained until convergence, (i) all features of the same class collapse to the corresponding class mean and (ii) the means associated with different classes are in a configuration where their pairwise distances are all equal and maximized. We extend such results and show through global solution and landscape analyses that a broad family of loss functions including commonly used label smoothing (LS) and focal loss (FL) exhibits Neural Collapse. Hence, all relevant losses (i.e., CE, LS, FL, MSE) produce equivalent features on training data. In particular, based on the unconstrained feature model assumption, we provide either the global landscape analysis for LS loss or the local landscape analysis for FL loss and show that the (only!) global minimizers are neural collapse solutions, while all other critical points are strict saddles whose Hessian exhibit negative curvature directions either in the global scope for LS loss or in the local scope for FL loss near the optimal solution. The experiments further show that Neural Collapse features obtained from all relevant losses (i.e., CE, LS, FL, MSE) lead to largely identical performance on test data as well, provided that the network is sufficiently large and trained until convergence.

## [LAMP: Extracting Text from Gradients with Language Model Priors](#)

- Mislav Balunovic · Dimitar Dimitrov · Nikola Jovanović · Martin Vechev
- abstract@[open-review](#): Recent work shows that sensitive user data can be reconstructed from gradient updates, breaking the key privacy promise of federated learning. While success was demonstrated primarily on image data, these methods do not directly transfer to other domains such as text. In this work, we propose LAMP, a novel attack tailored to textual data, that successfully reconstructs original text from gradients. Our key insight is to model the prior probability of the text with an auxiliary language model, utilizing it to guide the search towards more natural text. Concretely, LAMP introduces a discrete text transformation procedure that minimizes both the reconstruction loss and the prior text probability, as provided by the auxiliary language model. The procedure is alternated with a continuous optimization of the reconstruction loss, which also regularizes the length of the reconstructed embeddings. Our experiments demonstrate that LAMP reconstructs the original text significantly more precisely than prior work: we recover 5x more bigrams and 23% longer subsequences on average. Moreover, we are first to recover inputs from batch sizes larger than 1 for textual models. These findings indicate that gradient updates of models operating on textual data leak more information than previously thought.

## [Incrementality Bidding via Reinforcement Learning under Mixed and Delayed Rewards](#)

- Ashwinkumar Badanidiyuru Varadaraja · Zhe Feng · Tianxi Li · Haifeng Xu
- abstract@[open-review](#): Incrementality, which is used to measure the causal effect of showing an ad to a potential customer (e.g. a user in an internet platform) versus not, is a central object for advertisers in online advertising platforms. This paper investigates the problem of how an advertiser can learn to optimize the bidding sequence in an online manner without knowing the incrementality parameters in advance. We formulate the offline version of this problem as a specially structured episodic Markov Decision Process (MDP) and then, for its online learning counterpart, propose a novel reinforcement learning (RL) algorithm with regret at most  $\widetilde{O}(H^2\sqrt{T})$ , which depends on the number of rounds  $H$  and number of episodes  $T$ , but does not depend on the number of actions (i.e., possible bids). A fundamental difference between our learning problem from standard RL problems is that the realized reward feedback from conversion incrementality is mixed and delayed. To handle this difficulty we propose and analyze a novel pairwise moment-matching algorithm to learn the conversion incrementality, which we believe is of independent interest.

## [Efficient and Stable Fully Dynamic Facility Location](#)

- Sayan Bhattacharya · Silvio Lattanzi · Nikos Parotsidis
- abstract@[open-review](#): We consider the classic facility location problem in fully dynamic data streams, where elements can be both inserted and deleted. In this problem, one is interested in maintaining a stable and high quality solution throughout the data stream while using only little time per update (insertion or deletion). We study the problem and provide the first algorithm that at the same time maintains a constant approximation and incurs polylogarithmic amortized recourse per update. We complement our theoretical results with an experimental analysis showing the practical efficiency of our method.

## [UViM: A Unified Modeling Approach for Vision with Learned Guiding Codes](#)

- Alexander Kolesnikov · András Szabo Pinto · Lucas Beyer · Xiaohua Zhai · Jeremiah Harmsen · Neil Houlsby
- abstract@[open-review](#): We introduce UViM, a unified approach capable of modeling a wide range of computer vision tasks. In contrast to previous models, UViM has the same functional form for all tasks; it requires no task-specific modifications which require extensive human expertise. The approach involves two components: (I) a base model (feed-forward) which is trained to directly predict raw vision outputs, guided by a learned discrete code and (II) a language model (autoregressive) that is trained to generate the guiding code. These components complement each other: the language model is well-suited to modeling structured interdependent data, while the base model is efficient at dealing with high-dimensional outputs. We demonstrate the effectiveness of UViM on three diverse and challenging vision tasks: panoptic segmentation, depth prediction and image colorization, where we achieve competitive and near state-of-the-art results. Our experimental results suggest that UViM is a promising candidate for a unified modeling approach in computer vision.

## [Finding Correlated Equilibrium of Constrained Markov Game: A Primal-Dual Approach](#)

- Ziyi Chen · Shaocong Ma · Yi Zhou
- abstract@[open-review](#): Constrained Markov game is a fundamental problem that covers many applications, where multiple players compete with each other under behavioral constraints. The existing literature has proved the existence of Nash equilibrium for constrained Markov games, which turns out to be PPAD-complete and cannot be computed in polynomial time. In this work, we propose a surrogate notion of correlated equilibrium (CE) for constrained Markov games that can be computed in polynomial time, and study its fundamental properties. We show that the modification structure of CE of constrained Markov games is fundamentally different from that of unconstrained Markov games. Moreover, we prove that the corresponding Lagrangian function has zero duality gap. Based on this result, we develop the first primal-dual algorithm that provably converges to CE of constrained Markov games. In particular, we prove that both the duality gap and the constraint violation of the output policy converge at the rate  $\mathcal{O}(\frac{1}{\sqrt{T}})$ . Moreover, when adopting the V-learning algorithm as the subroutine in the primal update, our algorithm achieves an approximate CE with  $\epsilon$  duality gap with the sample complexity  $\mathcal{O}(H^9 \mathcal{S} \|\mathcal{A}\|^2 \epsilon^{-4})$ .

## [Fair Bayes-Optimal Classifiers Under Predictive Parity](#)

- Xianli Zeng · Edgar Dobriban · Guang Cheng
- abstract@[open-review](#): Increasing concerns about disparate effects of AI have motivated a great deal of work on fair machine learning. Existing works mainly focus on independence- and separation-based measures (e.g., demographic parity, equality of opportunity, equalized odds), while sufficiency-based measures such as predictive parity are much less studied. This paper considers predictive parity, which requires equalizing the probability of success given a positive prediction among different protected groups. We prove that, if the overall performances of different groups vary only moderately, all fair Bayes-optimal classifiers under predictive parity are group-wise thresholding rules. Perhaps surprisingly, this may not hold if group performance levels vary widely; in this case, we find that predictive parity among protected groups may lead to within-group unfairness. We then propose an algorithm we call FairBayes-DPP, aiming to ensure predictive parity when our condition is satisfied. FairBayes-DPP is an adaptive thresholding algorithm that aims to achieve predictive parity, while also seeking to maximize test accuracy. We provide supporting experiments conducted on synthetic and empirical data.

## [Rapid Model Architecture Adaption for Meta-Learning](#)

- Yiren Zhao · Xitong Gao · I Shumailov · Nicolo Fusi · Robert Mullins
- abstract@[open-review](#): Network Architecture Search (NAS) methods have recently gathered much attention. They design networks with better performance and use a much shorter search time compared to traditional manual tuning. Despite their efficiency in model deployments, most NAS algorithms target a single task on a fixed hardware system. However, real-life few-shot learning environments often cover a great number of tasks ( $T$ ) and deployments on a wide variety of hardware platforms ( $H$ ). The combinatorial search complexity  $T \times H$  creates a fundamental search efficiency challenge if one naively applies existing NAS methods to these scenarios. To overcome this issue, we show, for the first time, how to rapidly adapt model architectures to new tasks in a many-task many-hardware few-shot learning setup by integrating Model Agnostic Meta Learning (MAML) into the NAS flow. The proposed NAS method (H-Meta-NAS) is hardware-aware and performs optimisation in the MAML framework. MetaNAS shows a Pareto dominance compared to a variety of NAS and manual baselines in popular few-shot learning benchmarks with various hardware platforms and constraints. In particular, on the 5-way 1-shot Mini-ImageNet classification task, the proposed method outperforms the best manual baseline by a large margin (5.21% in accuracy) using 60% less computation.

## [LDSA: Learning Dynamic Subtask Assignment in Cooperative Multi-Agent Reinforcement Learning](#)

- Mingyu Yang · Jian Zhao · Xunhan Hu · Wengang Zhou · Jiangcheng Zhu · Houqiang Li
- abstract@[open-review](#): Cooperative multi-agent reinforcement learning (MARL) has made prominent progress in recent years. For training efficiency and scalability, most of the MARL algorithms make all agents share the same policy or value network. However, in many complex multi-agent tasks, different agents are expected to possess specific abilities to handle different subtasks. In those scenarios, sharing parameters indiscriminately may lead to similar behavior across all agents, which will limit the exploration efficiency and degrade the final performance. To balance the training complexity and the diversity of agent behavior, we propose a novel framework to learn dynamic subtask assignment (LDSA) in cooperative MARL. Specifically, we first introduce a subtask encoder to construct a vector representation for each subtask according to its identity. To reasonably assign agents to different subtasks, we propose an ability-based subtask selection strategy, which can dynamically group agents with similar abilities into the same subtask. In this way, agents dealing with the same subtask share their learning of specific abilities and different subtasks correspond to different specific abilities. We further introduce two regularizers to increase the representation difference between subtasks and stabilize the training by discouraging agents from frequently changing subtasks, respectively. Empirical results show that LDSA learns reasonable and effective subtask assignment for better collaboration and significantly improves the learning performance on the challenging StarCraft II micromanagement benchmark and Google Research Football.

## [SNN-RAT: Robustness-enhanced Spiking Neural Network through Regularized Adversarial Training](#)

- Jianhao Ding · Tong Bu · Zhaofei Yu · Jian Liu · Tiejun Huang
- abstract@[open-review](#): Spiking neural networks (SNNs) are promising to be widely deployed in real-time and safety-critical applications with the advance of neuromorphic computing. Recent work has demonstrated the insensitivity of SNNs to small random perturbations due to the discrete internal information representation. The variety of training algorithms and the involvement of the temporal dimension pose more threats to the robustness of SNNs than that of typical neural networks. We account for the vulnerability of SNNs by constructing adversaries based on differentiable approximation techniques. By deriving a Lipschitz constant specifically for the spike representation, we first theoretically answer the question of how much adversarial invulnerability is retained in SNNs. Hence, to defend against the broad attack methods, we propose a regularized adversarial training scheme with low computational overheads. SNNs can benefit from the constraint of the perturbed spike distance's amplification and the generalization on multiple adversarial  $\epsilon$ -neighbourhoods. Our experiments on the image recognition benchmarks have proven that our training scheme can defend against powerful adversarial attacks crafted from strong differentiable approximations. To be specific, our approach makes the black-box attacks of the Projected Gradient Descent attack nearly ineffective. We believe that our work will facilitate the spread of SNNs for safety-critical applications and help understand the robustness of the human brain.

## [Alternating Mirror Descent for Constrained Min-Max Games](#)

- Andre Wibisono · Molei Tao · Georgios Piliouras
- abstract@[open-review](#): In this paper we study two-player bilinear zero-sum games with constrained strategy spaces. An instance of natural occurrences of such constraints is when mixed strategies are used, which correspond to a probability simplex constraint. We propose and analyze the alternating mirror descent algorithm, in which each player takes turns to take action following the mirror descent algorithm for constrained optimization. We interpret alternating mirror descent as an alternating discretization of a skew-gradient flow in the dual space, and use tools from convex optimization and modified energy function to establish an  $\mathcal{O}(K^{-2/3})$  bound on its average regret after  $K$  iterations. This quantitatively verifies the algorithm's better behavior than the simultaneous version of mirror descent algorithm, which is known to diverge and yields an  $\mathcal{O}(K^{-1/2})$  average regret bound. In the special case of an unconstrained setting, our results recover the behavior of alternating gradient descent algorithm for zero-sum games which was studied in (Bailey et al., COLT 2020).

## [Graph Reordering for Cache-Efficient Near Neighbor Search](#)

- Benjamin Coleman · Santiago Segarra · Alexander Smola · Anshumali Shrivastava
- abstract@[open-review](#): Graph search is one of the most successful algorithmic trends in near neighbor search. Several of the most popular and empirically successful algorithms are, at their core, a greedy walk along a pruned near neighbor graph. However, graph traversal applications often suffer from poor memory access patterns, and near neighbor search is no exception to this rule. Our measurements show that popular search indices such as the hierarchical navigable small-world graph (HNSW) can have poor cache miss performance. To address this issue, we formulate the graph traversal problem as a cache hit maximization task and propose multiple graph reordering as a solution. Graph reordering is a memory layout optimization that groups commonly-accessed nodes together in memory. We mathematically formalize the connection between the graph layout and the cache complexity of search. We present exhaustive experiments applying several reordering algorithms to a leading graph-based near neighbor method based on the HNSW index. We find that reordering improves the query time by up to 40%, we present analysis and improvements for existing graph layout methods, and we demonstrate that the time needed to reorder the graph is negligible compared to the time required to construct the index.

## [CyCLIP: Cyclic Contrastive Language-Image Pretraining](#)

- Shashank Goel · Hritik Bansal · Sumit Bhatia · Ryan Rossi · Vishwa Vinay · Aditya Grover
- abstract@[open-review](#): Recent advances in contrastive representation learning over paired image-text data have led to models such as CLIP that achieve state-of-the-art performance for zero-shot classification and distributional robustness. Such models typically require joint reasoning in the image and text representation spaces for downstream inference tasks. Contrary to prior beliefs, we demonstrate that the image and text representations learned via a standard contrastive objective are not interchangeable and can lead to inconsistent downstream predictions. To mitigate this issue, we formalize consistency and propose CyCLIP, a framework for contrastive representation learning that explicitly optimizes for the learned representations to be geometrically consistent in the image and text space. In particular, we show that consistent representations can be learned by explicitly symmetrizing (a) the similarity between the two mismatched image-text pairs (cross-modal consistency); and (b) the similarity between the image-image pair and the text-text pair (in-modal consistency). Empirically, we show that the improved consistency in CyCLIP translates to significant gains over CLIP, with gains ranging from 10%-24% for zero-shot classification on standard benchmarks (CIFAR-10, CIFAR-100, ImageNet1K) and 10%-27% for robustness to various natural distribution shifts.

## [Towards a Standardised Performance Evaluation Protocol for Cooperative MARL](#)

- Rihab Gorsane · Oumayma Mahjoub · Ruan John de Kock · Roland Dubb · Siddarth Singh · Arnu Pretorius
- abstract@[open-review](#): Multi-agent reinforcement learning (MARL) has emerged as a useful approach to solving decentralised decision-making problems at scale. Research in the field has been growing steadily with many breakthrough algorithms proposed in recent years. In this work, we take a closer look at this rapid development with a focus on evaluation methodologies employed across a large body of research in cooperative MARL. By conducting a detailed meta-analysis of prior work, spanning 75 papers accepted for publication from 2016 to 2022, we bring to light worrying trends that put into question the true rate of progress. We further consider these trends in a wider context and take inspiration from single-agent RL literature on similar issues with recommendations that remain applicable to MARL. Combining these recommendations, with novel insights from our analysis, we propose a standardised performance evaluation protocol for cooperative MARL. We argue that such a standard protocol, if widely adopted, would greatly improve the validity and credibility of future research, make replication and reproducibility easier, as well as improve the ability of the field to accurately gauge the rate of progress over time by being able to make sound comparisons across different works. Finally, we release our meta-analysis data publicly for future research on evaluation.

## [GREED: A Neural Framework for Learning Graph Distance Functions](#)

- Rishabh Ranjan · Siddharth Grover · Sourav Medya · Venkatesan Chakaravarthy · Yogish Sabharwal · Sayan Ranu
- abstract@[open-review](#): Similarity search in graph databases is one of the most fundamental operations in graph analytics. Among various distance functions, graph and subgraph edit distances (GED and SED respectively) are two of the most popular and expressive measures. Unfortunately, exact computations for both are NP-hard. To overcome this computational bottleneck, neural approaches to learn and predict edit distance in polynomial time have received much interest. While considerable progress has been made, there exist limitations that need to be addressed. First, the efficacy of an approximate distance function lies not only in its approximation accuracy, but also in the preservation of its properties. To elaborate, although GED is a metric, its neural approximations do not provide such a guarantee. This prohibits their usage in higher order tasks that rely on metric distance functions, such as clustering or indexing. Second, several existing frameworks for GED do not extend to SED due to SED being asymmetric. In this work, we design a novel siamese graph neural network called Greed, which through a carefully crafted inductive bias, learns GED and SED in a property-preserving manner. Through extensive experiments across \$10\$ real graph datasets containing up to \$7\$ million edges, we establish that Greed is not only more accurate than the state of the art, but also up to \$3\$ orders of magnitude faster. Even more significantly, due to preserving the triangle inequality, the generated embeddings are indexable and consequently, even in a CPU-only environment, Greed is up to \$50\$ times faster than GPU-powered computations of the closest baseline.

## [Coarse-to-Fine Vision-Language Pre-training with Fusion in the Backbone](#)

- Zi-Yi Dou · Aishwarya Kamath · Zhe Gan · Pengchuan Zhang · Jianfeng Wang · Linjie Li · Zicheng Liu · Ce Liu · Yann LeCun · Nanyun Peng · Jianfeng Gao · Lijuan Wang
- abstract@[open-review](#): Vision-language (VL) pre-training has recently received considerable attention. However, most existing end-to-end pre-training approaches either only aim to tackle VL tasks such as image-text retrieval, visual question answering (VQA) and image captioning that test high-level understanding of images, or only target region-level understanding for tasks such as phrase grounding and object detection. We present FIBER (Fusion-In-the-Backbone-based transformER), a new VL model architecture that can seamlessly handle both these types of tasks. Instead of having dedicated transformer layers for fusion after the uni-modal backbones, FIBER pushes multimodal fusion deep into the model by inserting cross-attention into the image and text backbones to better capture multimodal interactions. In addition, unlike previous work that is either only pre-trained on image-text data or on fine-grained data with box-level annotations, we present a two-stage pre-training strategy that uses both these kinds of data efficiently: (i) coarse-grained pre-training based on image-text data; followed by (ii) fine-grained pre-training based on image-text-box data. We conduct comprehensive experiments on a wide range of VL tasks, ranging from VQA, image captioning, and retrieval, to phrase grounding, referring expression comprehension, and object detection. Using deep multimodal fusion coupled with the two-stage pre-training, FIBER provides consistent performance improvements over strong baselines across all tasks, often outperforming methods using magnitudes more data. Code will be released upon acceptance.

## [Spectral Bias Outside the Training Set for Deep Networks in the Kernel Regime](#)

- Benjamin Bowman · Guido Montufar
- abstract@[open-review](#): We provide quantitative bounds measuring the  $L^2$  difference in function space between the trajectory of a finite-width network trained on finitely many samples from the idealized kernel dynamics of infinite width and infinite data. An implication of the bounds is that the network is biased to learn the top eigenfunctions of the Neural Tangent Kernel not just on the training set but over the entire input space. This bias depends on the model architecture and input distribution alone and thus does not depend on the target function which does not need to be in the RKHS of the kernel. The result is valid for deep architectures with fully connected, convolutional, and residual layers. Furthermore the width does not need to grow polynomially with the number of samples in order to obtain high probability bounds up to a stopping time. The proof exploits the low-effective-rank property of the Fisher Information Matrix at initialization, which implies a low effective dimension of the model (far smaller than the number of parameters). We conclude that local capacity control from the low effective rank of the Fisher Information Matrix is still underexplored theoretically.

## Provably Efficient Reinforcement Learning in Partially Observable Dynamical Systems

- Masatoshi Uehara · Ayush Sekhari · Jason Lee · Nathan Kallus · Wen Sun
- abstract@[open-review](#): We study Reinforcement Learning for partially observable systems using function approximation. We propose a new PO-bilinear framework, that is general enough to include models such as undercomplete tabular Partially Observable Markov Decision Processes (POMDPs), Linear Quadratic Gaussian (LQG), Predictive State Representations (PSRs), as well as a newly introduced model Hilbert Space Embeddings of POMDPs. Under this framework, we propose an actor-critic style algorithm that is capable to performing agnostic policy learning. Given a policy class that consists of memory based policies (i.e., policy that looks at a fixed-length window of recent observations), and a value function class that consists of functions taking both memory and future observations as inputs, our algorithm learns to compete against the best memory-based policy among the policy class. For certain examples such as undercomplete POMDPs and LQGs, by leveraging their special properties, our algorithm is even capable of competing against the globally optimal policy without paying an exponential dependence on the horizon.

## Adapting to Online Label Shift with Provable Guarantees

- Yong Bai · Yu-Jie Zhang · Peng Zhao · Masashi Sugiyama · Zhi-Hua Zhou
- abstract@[open-review](#): The standard supervised learning paradigm works effectively when training data shares the same distribution as the upcoming testing samples. However, this assumption is often violated in real-world applications, especially when testing data appear in an online fashion. In this paper, we formulate and investigate the problem of \emph{online label shift} (OLS): the learner trains an initial model from the labeled offline data and then deploys it to an unlabeled online environment where the underlying label distribution changes over time but the label-conditional density does not. The non-stationarity nature and the lack of supervision make the problem challenging to be tackled. To address the difficulty, we construct a new unbiased risk estimator that utilizes the unlabeled data, which exhibits many benign properties albeit with potential non-convexity. Building upon that, we propose novel online ensemble algorithms to deal with the non-stationarity of the environments. Our approach enjoys optimal \emph{dynamic regret}, indicating that the performance is competitive with a clairvoyant who knows the online environments in hindsight and then chooses the best decision for each round. The obtained dynamic regret bound scales with the intensity and pattern of label distribution shift, hence exhibiting the adaptivity in the OLS problem. Extensive experiments are conducted to validate the effectiveness and support our theoretical findings.

## A Classification of $\$G\$$ -invariant Shallow Neural Networks

- Devanshu Agrawal · James Ostrowski
- abstract@[open-review](#): When trying to fit a deep neural network (DNN) to a  $\$G\$$ -invariant target function with respect to a group  $\$G\$$ , it only makes sense to constrain the DNN to be  $\$G\$$ -invariant as well. However, there can be many different ways to do this, thus raising the problem of " $\$G\$$ -invariant neural architecture design": What is the optimal  $\$G\$$ -invariant architecture for a given problem? Before we can consider the optimization problem itself, we must understand the search space, the architectures in it, and how they relate to one another. In this paper, we take a first step towards this goal; we prove a theorem that gives a classification of all  $\$G\$$ -invariant single-hidden-layer or "shallow" neural network ( $\$G\$$ -SNN) architectures with ReLU activation for any finite orthogonal group  $\$G\$$ . The proof is based on a correspondence of every  $\$G\$$ -SNN to a signed permutation representation of  $\$G\$$  acting on the hidden neurons. The classification is equivalently given in terms of the first cohomology classes of  $\$G\$$ , thus admitting a topological interpretation. Based on a code implementation, we enumerate the  $\$G\$$ -SNN architectures for some example groups  $\$G\$$  and visualize their structure. We draw the network morphisms between the enumerated architectures that can be leveraged during neural architecture search (NAS). Finally, we prove that architectures corresponding to inequivalent cohomology classes in a given cohomology ring coincide in function space only when their weight matrices are zero, and we discuss the implications of this in the context of NAS.

## When to Ask for Help: Proactive Interventions in Autonomous Reinforcement Learning

- Annie Xie · Fahim Tajwar · Archit Sharma · Chelsea Finn
- abstract@[open-review](#): A long-term goal of reinforcement learning is to design agents that can autonomously interact and learn in the world. A critical challenge to such autonomy is the presence of irreversible states which require external assistance to recover from, such as when a robot arm has pushed an object off of a table. While standard agents require constant monitoring to decide when to intervene, we aim to design proactive agents that can request human intervention only when needed. To this end, we propose an algorithm that can efficiently learn to detect and avoid states that are irreversible, and proactively ask for help in case the agent does enter them. On a suite of continuous control environments with unknown irreversible states, we find that our algorithm exhibits both better sample- and intervention-efficiency compared to existing methods.

## A Damped Newton Method Achieves Global $\mathcal{O}(\frac{1}{k^2})$ and Local Quadratic Convergence Rate

- Slavomír Hanzely · Dmitry Kamzolov · Dmitry Pasechnyuk · Alexander Gasnikov · Peter Richtarik · Martin Takac
- abstract@[open-review](#): In this paper, we present the first stepsize schedule for Newton method resulting in fast global and local convergence guarantees. In particular, we a) prove an  $\mathcal{O}(1/k^2)$  global rate, which matches the state-of-the-art global rate of cubically regularized Newton method of Polyak and Nesterov (2006) and of regularized Newton method of Mishchenko (2021), and the later variant of Doikov and Nesterov (2021), b) prove a local quadratic rate, which matches the best-known local rate of second-order methods, and c) our stepsize formula is simple, explicit, and does not require solving any subproblem. Our convergence proofs hold under affine-invariant assumptions closely related to the notion of self-concordance. Finally, our method has competitive performance when compared to existing baselines which share the same fast global convergence guarantees.

## Mind the Gap: Understanding the Modality Gap in Multi-modal Contrastive Representation Learning

- Victor Weixin Liang · Yuhui Zhang · Yongchan Kwon · Serena Yeung · James Zou
- abstract@[open-review](#): We present modality gap, an intriguing geometric phenomenon of the representation space of multi-modal models. Specifically, we show that different data modalities (e.g. images and text) are embedded at arm's length in their shared representation in multi-modal models such as CLIP. Our systematic analysis demonstrates that this gap is caused by a combination of model initialization and contrastive learning optimization. In model initialization, we show empirically and theoretically that the representation of a common deep neural network is restricted to a narrow cone. As a consequence, in a multi-modal model with two encoders, the representations of the two modalities are clearly apart when the model is initialized. During optimization, contrastive learning keeps the different modalities separate by a certain distance, which is influenced by the temperature parameter in the loss function. Our experiments further demonstrate that varying the modality gap distance has a significant impact in improving the model's downstream zero-shot classification performance and fairness.

## Trajectory balance: Improved credit assignment in GFlowNets

- Nikolay Malkin · Moksh Jain · Emmanuel Bengio · Chen Sun · Yoshua Bengio
- abstract@[open-review](#): Generative Flow Networks (GFlowNets) are a method for learning a stochastic policy for generating compositional objects, such as graphs or strings, from a given unnormalized density by sequences of actions, where many possible action sequences may lead to the same object. We find previously proposed learning objectives for GFlowNets, flow matching and detailed balance, which are analogous to temporal difference learning, to be prone to inefficient credit propagation across long action sequences. We thus propose a new learning objective for GFlowNets, trajectory balance, as a more efficient alternative to previously used objectives. We prove that any global minimizer of the trajectory balance objective can define a policy that

samples exactly from the target distribution. In experiments on four distinct domains, we empirically demonstrate the benefits of the trajectory balance objective for GFlowNet convergence, diversity of generated samples, and robustness to long action sequences and large action spaces.

## [Supervised Training of Conditional Monge Maps](#)

- Charlotte Bunne · Andreas Krause · Marco Cuturi
- abstract@[open-review](#): Optimal transport (OT) theory describes general principles to define and select, among many possible choices, the most efficient way to map a probability measure onto another. That theory has been mostly used to estimate, given a pair of source and target probability measures  $(\mu, \nu)$ , a parameterized map  $T_\theta$  that can efficiently map  $\mu$  onto  $\nu$ . In many applications, such as predicting cell responses to treatments, the data measures  $\mu, \nu$  (features of untreated/treated cells) that define optimal transport problems do not arise in isolation but are associated with a context  $c$  (the treatment). To account for and incorporate that context in OT estimation, we introduce  $\text{CondOT}$ , an approach to estimate OT maps conditioned on a context variable, using several pairs of measures  $(\mu_i, \nu_i)$  tagged with a context label  $c_i$ . Our goal is to learn a global map  $T_\theta$  which is not only expected to fit all pairs in the dataset  $(c_i, (\mu_i, \nu_i))$ , i.e.,  $T_\theta(c_i) \approx \mu_i \approx \nu_i$ , but should generalize to produce meaningful maps  $T_\theta(c_{\text{new}})$  conditioned on unseen contexts  $c_{\text{new}}$ . Our approach harnesses and provides a novel usage for partially input convex neural networks, for which we introduce a robust and efficient initialization strategy inspired by Gaussian approximations. We demonstrate the ability of  $\text{CondOT}$  to infer the effect of an arbitrary combination of genetic or therapeutic perturbations on single cells, using only observations of the effects of said perturbations separately.

## [Efficient Risk-Averse Reinforcement Learning](#)

- Ido Greenberg · Yinlam Chow · Mohammad Ghavamzadeh · Shie Mannor
- abstract@[open-review](#): In risk-averse reinforcement learning (RL), the goal is to optimize some risk measure of the returns. A risk measure often focuses on the worst returns out of the agent's experience. As a result, standard methods for risk-averse RL often ignore high-return strategies. We prove that under certain conditions this inevitably leads to a local-optimum barrier, and propose a mechanism we call soft risk to bypass it. We also devise a novel cross entropy module for sampling, which (1) preserves risk aversion despite the soft risk; (2) independently improves sample efficiency. By separating the risk aversion of the sampler and the optimizer, we can sample episodes with poor conditions, yet optimize with respect to successful strategies. We combine these two concepts in CeSoR - Cross-entropy Soft-Risk optimization algorithm - which can be applied on top of any risk-averse policy gradient (PG) method. We demonstrate improved risk aversion in maze navigation, autonomous driving, and resource allocation benchmarks, including in scenarios where standard risk-averse PG completely fails.

## [Constants of motion network](#)

- Muhammad Firmansyah Kasim · Yi Heng Lim
- abstract@[open-review](#): The beauty of physics is that there is usually a conserved quantity in an always-changing system, known as the constant of motion. Finding the constant of motion is important in understanding the dynamics of the system, but typically requires mathematical proficiency and manual analytical work. In this paper, we present a neural network that can simultaneously learn the dynamics of the system and the constants of motion from data. By exploiting the discovered constants of motion, it can produce better predictions on dynamics and can work on a wider range of systems than Hamiltonian-based neural networks. In addition, the training progresses of our method can be used as an indication of the number of constants of motion in a system which could be useful in studying a novel physical system.

## [An Invisible Issue of Task Underspecification in Deep Reinforcement Learning](#)

- Vindula Jayawardana · Catherine Tang · Sirui Li · Dajiang Suo · Cathy Wu
- abstract@[open-review](#): Performance evaluations of Deep Reinforcement Learning (DRL) algorithms are an integral part of the scientific progress of the field. However, standard performance evaluation practices in evaluating algorithmic generalization within a task of DRL methods can be unreliable and misleading if not careful. An important source of possible error lies in the reliance of the reported outcomes on often arbitrarily selected point Markov decision processes (point MDPs), stemming from task underspecification. A large class of DRL tasks, particularly in real-world decision problems, induce a family of MDPs, which---perhaps confusingly---each has the same high-level problem definition. As a demonstrative example, consider that a classic pendulum control task could be represented by a family of possible MDPs, each with a different pendulum mass, but is typically represented as a single MDP. This article argues that for reliable downstream decision-making, performance evaluations on a task in DRL should be carried out over a family of MDPs rather than a point MDP, which may be subject to bias. This article first illustrates the pitfalls of point MDP based evaluations through benchmark DRL control tasks and a real-world case study in traffic signal control. Then, significant inconsistencies between conclusions derived from point MDP based evaluations and MDP family-based evaluations are presented. Subsequently, to overcome the prohibitive cost of training DRL models on entire families of MDPs, a series of recommendations is provided to perform accurate yet efficient performance evaluations under a computational budget. This work contributes to bolstering the empirical rigor of reinforcement learning, especially as the outcomes of DRL trickle into downstream decision-making in real-world contexts.

## [Supervising the Multi-Fidelity Race of Hyperparameter Configurations](#)

- Martin Wistuba · Arlind Kadra · Josif Grabocka
- abstract@[open-review](#): Multi-fidelity (gray-box) hyperparameter optimization techniques (HPO) have recently emerged as a promising direction for tuning Deep Learning methods. However, existing methods suffer from a sub-optimal allocation of the HPO budget to the hyperparameter configurations. In this work, we introduce DyHPO, a Bayesian Optimization method that learns to decide which hyperparameter configuration to train further in a dynamic race among all feasible configurations. We propose a new deep kernel for Gaussian Processes that embeds the learning curve dynamics, and an acquisition function that incorporates multi-budget information. We demonstrate the significant superiority of DyHPO against state-of-the-art hyperparameter optimization methods through large-scale experiments comprising 50 datasets (Tabular, Image, NLP) and diverse architectures (MLP, CNN/NAS, RNN).

## [Implicit Bias of Gradient Descent on Reparametrized Models: On Equivalence to Mirror Descent](#)

- Zhiyuan Li · Tianhao Wang · Jason Lee · Sanjeev Arora
- abstract@[open-review](#): As part of the effort to understand implicit bias of gradient descent in overparametrized models, several results have shown how the training trajectory on the overparametrized model can be understood as mirror descent on a different objective. The main result here is a complete characterization of this phenomenon under a notion termed commuting parametrization, which encompasses all the previous results in this setting. It is shown that gradient flow with any commuting parametrization is equivalent to continuous mirror descent with a related mirror map. Conversely, continuous mirror descent with any mirror map can be viewed as gradient flow with a related commuting parametrization. The latter result relies upon Nash's embedding theorem.

## [Learning to Branch with Tree MDPs](#)

- Lara Scavuzzo · Feng Chen · Didier Chetelat · Maxime Gasse · Andrea Lodi · Neil Yorke-Smith · Karen Aardal
- abstract@[open-review](#): State-of-the-art Mixed Integer Linear Programming (MILP) solvers combine systematic tree search with a plethora of hard-coded heuristics, such as branching rules. While approaches to learn branching strategies have received increasing attention and have shown very promising results, most of the literature focuses on learning fast approximations of the `strong branching` rule. Instead, we propose to learn branching rules from scratch with Reinforcement Learning (RL). We revisit the work of Etheve et al. (2020) and propose a generalization of Markov Decisions Processes (MDP), which we call `tree MDP`, that provides a more suitable formulation of the branching problem. We derive a policy gradient theorem for tree MDPs that exhibits a better credit assignment compared to its temporal counterpart. We demonstrate through computational experiments that this new framework is suitable to tackle the learning-to-branch problem in MILP, and improves the learning convergence.

## [Adaptive Bio-Inspired Fish Simulation with Deep Reinforcement Learning](#)

- Yuko Ishiwaka · Xiao Zeng · Shun Ogawa · Donovan Westwater · Tadayuki Tone · Masaki Nakada
- abstract@[open-review](#): Our goal is to synthesize realistic underwater scenes with various fish species in different fish cages, which can be utilized to train computer vision models to automate fish counting and sizing tasks. It is a challenging problem to prepare a sufficiently diverse labeled dataset of images from aquatic environments. We solve this challenge by introducing an adaptive bio-inspired fish simulation. The behavior of caged fish changes based on the species, size and number of fish, and the size and shape of the cage, among other variables. However, a method to autonomously achieve schooling behavior for caged fish did not exist. In this paper, we propose a method for achieving schooling behavior for any given combination of variables, using multi-agent deep reinforcement learning (DRL) in various fish cages in arbitrary environments. Furthermore, to visually reproduce the underwater scene in different locations and seasons, we incorporate a physically-based underwater simulation.

## [Marksman Backdoor: Backdoor Attacks with Arbitrary Target Class](#)

- Khoa D Doan · Yingjie Lao · Ping Li
- abstract@[open-review](#): In recent years, machine learning models have been shown to be vulnerable to backdoor attacks. Under such attacks, an adversary embeds a stealthy backdoor into the trained model such that the compromised models will behave normally on clean inputs but will misclassify according to the adversary's control on maliciously constructed input with a trigger. While these existing attacks are very effective, the adversary's capability is limited: given an input, these attacks can only cause the model to misclassify toward a single pre-defined or target class. In contrast, this paper exploits a novel backdoor attack with a much more powerful payload, denoted as Marksman, where the adversary can arbitrarily choose which target class the model will misclassify given any input during inference. To achieve this goal, we propose to represent the trigger function as a class-conditional generative model and to inject the backdoor in a constrained optimization framework, where the trigger function learns to generate an optimal trigger pattern to attack any target class at will while simultaneously embedding this generative backdoor into the trained model. Given the learned trigger-generation function, during inference, the adversary can specify an arbitrary backdoor attack target class, and an appropriate trigger causing the model to classify toward this target class is created accordingly. We show empirically that the proposed framework achieves high attack performance (e.g., 100% attack success rates in several experiments) while preserving the clean-data performance in several benchmark datasets, including MNIST, CIFAR10, GTSRB, and TinyImageNet. The proposed Marksman backdoor attack can also easily bypass existing backdoor defenses that were originally designed against backdoor attacks with a single target class. Our work takes another significant step toward understanding the extensive risks of backdoor attacks in practice.

## [Towards Understanding Grokking: An Effective Theory of Representation Learning](#)

- Ziming Liu · Ouail Kitouni · Niklas S Nolte · Eric Michaud · Max Tegmark · Mike Williams
- abstract@[open-review](#): We aim to understand grokking, a phenomenon where models generalize long after overfitting their training set. We present both a microscopic analysis anchored by an effective theory and a macroscopic analysis of phase diagrams describing learning performance across hyperparameters. We find that generalization originates from structured representations, whose training dynamics and dependence on training set size can be predicted by our effective theory (in a toy setting). We observe empirically the presence of four learning phases: comprehension, grokking, memorization, and confusion. We find representation learning to occur only in a "Goldilocks zone" (including comprehension and grokking) between memorization and confusion. Compared to the comprehension phase, the grokking phase stays closer to the memorization phase, leading to delayed generalization. The Goldilocks phase is reminiscent of "intelligence from starvation" in Darwinian evolution, where resource limitations drive discovery of more efficient solutions. This study not only provides intuitive explanations of the origin of grokking, but also highlights the usefulness of physics-inspired tools, e.g., effective theories and phase diagrams, for understanding deep learning.

## [Reinforcement Learning with Logarithmic Regret and Policy Switches](#)

- Grigoris Velekas · Zhuoran Yang · Amin Karbasi
- abstract@[open-review](#): In this paper, we study the problem of regret minimization for episodic Reinforcement Learning (RL) both in the model-free and the model-based setting. We focus on learning with general function classes and general model classes, and we derive results that scale with the eluder dimension of these classes. In contrast to the existing body of work that mainly establishes instance-independent regret guarantees, we focus on the instance-dependent setting and show that the regret scales logarithmically with the horizon  $T$ , provided that there is a gap between the best and the second best action in every state. In addition, we show that such a logarithmic regret bound is realizable by algorithms with  $O(\log T)$  switching cost (also known as adaptivity complexity). In other words, these algorithms rarely switch their policy during the course of their execution. Finally, we complement our results with lower bounds which show that even in the tabular setting, we cannot hope for regret guarantees lower than  $O(\log T)$ .

## [A Simple Approach to Automated Spectral Clustering](#)

- Jicong Fan · Yiheng Tu · Zhao Zhang · Mingbo Zhao · Haijun Zhang
- abstract@[open-review](#): The performance of spectral clustering heavily relies on the quality of affinity matrix. A variety of affinity-matrix-construction (AMC) methods have been proposed but they have hyperparameters to determine beforehand, which requires strong experience and leads to difficulty in real applications, especially when the inter-cluster similarity is high and/or the dataset is large. In addition, we often need to choose different AMC methods for different datasets, which still depends on experience. To solve these two challenging problems, in this paper, we present a simple yet effective method for automated spectral clustering. First, we propose to find the most reliable affinity matrix via grid search or Bayesian optimization among a set of candidates given by different AMC methods with different hyperparameters, where the reliability is quantified by the `relative-eigen-gap` of graph Laplacian introduced in this paper. Second, we propose a fast and accurate AMC method based on least squares representation and thresholding and prove its effectiveness theoretically. Finally, we provide a large-scale extension for the automated spectral clustering method, of which the time complexity is linear with the number of data points. Extensive experiments of natural image clustering show that our method is more versatile, accurate, and efficient than baseline methods.

## [SeqPATE: Differentially Private Text Generation via Knowledge Distillation](#)

- zhiliang tian · Yingxiu Zhao · Ziyue Huang · Yu-Xiang Wang · Nevin L. Zhang · He He
- abstract@[open-review](#): Protecting the privacy of user data is crucial for text generation models, which may leak sensitive information during generation. Differentially private (DP) learning methods provide guarantees against identifying the existence of a training sample from model outputs. PATE is a recent DP learning algorithm that achieves high utility with strong privacy protections on training samples. However, text generations conduct sequential

generations on a large output space, and PATE is not customized for that paradigm. Besides, PATE does well in protecting sample-level privacy but requires a high privacy cost on protecting phrases in samples. In this paper, we propose SeqPATE, an extension of PATE on text generation, which aims to protect the privacy of both training samples and sensitive phrases in samples. To adapt to text generations, we generate pseudo inputs and reduce the sequence generation problem to next word predictions. To handle the large output space, we propose a candidate filtering strategy to dynamically reduce the space, and refine the teacher aggregation of PATE to avoid low agreement due to voting for a large number of candidates. To reduce privacy losses, we design an efficient knowledge distillation to decrease the frequency of querying teachers. The experiments verify the effectiveness of SeqPATE in protecting both samples and sensitive phrases.

## [Sparse Hypergraph Community Detection Thresholds in Stochastic Block Model](#)

- Erchuan Zhang · David Suter · Giang Truong · Syed Zulqarnain Gilani
- abstract@[open-review](#): Community detection in random graphs or hypergraphs is an interesting fundamental problem in statistics, machine learning and computer vision. When the hypergraphs are generated by a {em stochastic block model}, the existence of a sharp threshold on the model parameters for community detection was conjectured by Angelini et al. 2015. In this paper, we confirm the positive part of the conjecture, the possibility of non-trivial reconstruction above the threshold, for the case of two blocks by comparing the hypergraph stochastic block model with its Erd{"o}s-R{\'e}nyi counterpart. Furthermore, we show the negative part of the conjecture by relating the model with the so-called {em multi-type Galton-Watson hypertrees} and considering the broadcasting problem on these hypertrees. The methods developed in this paper are generalised from the study of sparse random graphs by Mossel et al. 2015.

## [Confident Approximate Policy Iteration for Efficient Local Planning in \\$q^\pi\\$-realizable MDPs](#)

- Gell{\'e}rt Weisz · Andr{\'a}s Gy{\'o}rgy · Csaba Szepesv{\'a}ri
- abstract@[open-review](#): We consider approximate dynamic programming in \$\gamma\$-discounted Markov Decision Processes and apply it to approximate planning with linear value function approximation. Our first contribution is a new variant of Approximate Policy Iteration (API), which we call Confident Approximate Policy Iteration (CAPI). CAPI is shown to produce a deterministic stationary policy with an optimal error bound scaling linearly with product of the effective horizon \$H\$ and the worst-case approximation error \$\epsilon\$ of the action-value functions \$q^\pi\$ of stationary policies. This improvement over API (whose error scales with \$H^2\$) comes at the price of an \$H\$-fold increase in memory cost. Unlike Scherrer and Lesner [2012], who recommended returning a non-stationary policy to achieve a similar improvement (with the same memory overhead), we are able to stick to stationary policies. This allows for our second contribution, which is the application of CAPI to planning with local access to a simulator and \$d\$-dimensional linear function approximation. As such, we design a planning algorithm that applies CAPI to obtain a sequence of policies with successively refined accuracies on a dynamically evolving set of states. The algorithm outputs an \$O(\sqrt{d}H\epsilon)\$-optimal policy after issuing \$\tilde{O}(dH^4/\epsilon^2)\$ queries to the simulator, simultaneously achieving the best known bounds for both accuracy and query complexity, while earlier algorithms in the literature achieve only one of them. These improvements come at the expense of a mild (polynomial) increase in memory and computational costs of both the algorithm and the policy found by the algorithm.

## [Self-Explaining Deviations for Coordination](#)

- Hengyuan Hu · Samuel Sokota · David Wu · Anton Bakhtin · Andrei Lupu · Brandon Cui · Jakob Foerster
- abstract@[open-review](#): Fully cooperative, partially observable multi-agent problems are ubiquitous in the real world. In this paper, we focus on a specific subclass of coordination problems in which humans are able to discover self-explaining deviations (SEDs). SEDs are actions that deviate from the common understanding of what reasonable behavior would be in normal circumstances. They are taken with the intention of causing another agent or other agents to realize, using theory of mind, that the circumstance must be abnormal. We first motivate SED with a real world example and formalize its definition. Next, we introduce a novel algorithm, IMPROvement maxImizing Self-Explaining Deviations (IMPROVISED), to perform SEDs. Lastly, we evaluate IMPROVISED both in an illustrative toy setting and the popular benchmark setting Hanabi, where it is the first method to produce so called finesse plays, which are regarded as one of the more iconic examples of human theory of mind.

## [Planning to the Information Horizon of BAMDPs via Epistemic State Abstraction](#)

- Dilip Arumugam · Satinder Singh
- abstract@[open-review](#): The Bayes-Adaptive Markov Decision Process (BAMDP) formalism pursues the Bayes-optimal solution to the exploration-exploitation trade-off in reinforcement learning. As the computation of exact solutions to Bayesian reinforcement-learning problems is intractable, much of the literature has focused on developing suitable approximation algorithms. In this work, before diving into algorithm design, we first define, under mild structural assumptions, a complexity measure for BAMDP planning. As efficient exploration in BAMDPs hinges upon the judicious acquisition of information, our complexity measure highlights the worst-case difficulty of gathering information and exhausting epistemic uncertainty. To illustrate its significance, we establish a computationally-intractable, exact planning algorithm that takes advantage of this measure to show more efficient planning. We then conclude by introducing a specific form of state abstraction with the potential to reduce BAMDP complexity that gives rise to a computationally-tractable, approximate planning algorithm.

## [Neural Lyapunov Control of Unknown Nonlinear Systems with Stability Guarantees](#)

- Ruikun Zhou · Thanin Quartz · Hans De Sterck · Jun Liu
- abstract@[open-review](#): Learning for control of dynamical systems with formal guarantees remains a challenging task. This paper proposes a learning framework to simultaneously stabilize an unknown nonlinear system with a neural controller and learn a neural Lyapunov function to certify a region of attraction (ROA) for the closed-loop system with provable guarantees. The algorithmic structure consists of two neural networks and a satisfiability modulo theories (SMT) solver. The first neural network is responsible for learning the unknown dynamics. The second neural network aims to identify a valid Lyapunov function and a provably stabilizing nonlinear controller. The SMT solver verifies the candidate Lyapunov function satisfies the Lyapunov conditions. We further provide theoretical guarantees of the proposed learning framework and show that the obtained Lyapunov function indeed verifies for the unknown nonlinear system under mild assumptions. We illustrate the effectiveness of the results with a few numerical experiments.

## [Computationally Efficient Aggregated Kernel Tests using Incomplete \\$U\\$-statistics](#)

- Antonin Schrab · Ilmun Kim · Benjamin Guedj · Arthur Gretton
- abstract@[open-review](#): We propose a series of computationally efficient, nonparametric tests for the two-sample, independence and goodness-of-fit problems, using the Maximum Mean Discrepancy (MMD), Hilbert Schmidt Independence Criterion (HSIC), and Kernel Stein Discrepancy (KSD), respectively. Our test statistics are incomplete \$U\$-statistics, with a computational cost that interpolates between linear time in the number of samples, and quadratic time, as associated with classical \$U\$-statistic tests. The three proposed tests aggregate over several kernel bandwidths to detect departures from the null on various scales: we call the resulting tests MMDAggInc, HSICAggInc and KSDAggInc. For the test thresholds, we derive a quantile bound for wild bootstrapped incomplete \$U\$-statistics, which is of independent interest. We derive uniform separation rates for MMDAggInc and HSICAggInc, and quantify exactly the trade-off between computational efficiency and the attainable rates: this result is novel for tests based on incomplete \$U\$-statistics, to our knowledge. We further show that in the quadratic-time case, the wild bootstrap incurs no penalty to test power over more widespread permutation-based approaches, since both attain the same minimax optimal rates (which in turn match the rates that use oracle quantiles). We support our claims with numerical experiments on the trade-off between computational efficiency and test power.

## [Learning Enhanced Representation for Tabular Data via Neighborhood Propagation](#)

- Kounianhua Du · Weinan Zhang · Ruiwen Zhou · Yangkun Wang · Xilong Zhao · Jiarui Jin · Quan Gan · Zheng Zhang · David P Wipf
- abstract@[open-review](#): Prediction over tabular data is an essential and fundamental problem in many important downstream tasks. However, existing methods either take a data instance of the table independently as input or do not fully utilize the multi-row features and labels to directly change and enhance the target data representations. In this paper, we propose to 1) construct a hypergraph from relevant data instance retrieval to model the cross-row and cross-column patterns of those instances, and 2) perform message Propagation to Enhance the target data instance representation for Tabular prediction tasks. Specifically, our specially-designed message propagation step benefits from 1) the fusion of label and features during propagation, and 2) locality-aware multiplicative high-order interaction between features. Experiments on two important tabular prediction tasks validate the superiority of the proposed PET model against other baselines. Additionally, we demonstrate the effectiveness of the model components and the feature enhancement ability of PET via various ablation studies and visualizations.

## [DigGAN: Discriminator gradient Gap Regularization for GAN Training with Limited Data](#)

- Tiantian Fang · Ruoyu Sun · Alex Schwing
- abstract@[open-review](#): Generative adversarial nets (GANs) have been remarkably successful at learning to sample from distributions specified by a given dataset, particularly if the given dataset is reasonably large compared to its dimensionality. However, given limited data, classical GANs have struggled, and strategies like output-regularization, data-augmentation, use of pre-trained models and pruning have been shown to lead to improvements. Notably, the applicability of these strategies is often constrained to particular settings, e.g., availability of a pretrained GAN, or increases training time, e.g., when using pruning. In contrast, we propose a Discriminator gradient Gap regularized GAN (DigGAN) formulation which can be added to any existing GAN. DigGAN augments existing GANs by encouraging to narrow the gap between the norm of the gradient of a discriminator's prediction w.r.t. real images and w.r.t. the generated samples. We observe this formulation to avoid bad attractors within the GAN loss landscape, and we find DigGAN to significantly improve the results of GAN training when limited data is available.

## [On the Interpretability of Regularisation for Neural Networks Through Model Gradient Similarity](#)

- Vincent Szolnoky · Viktor Andersson · Balazs Kulcsar · Rebecka Jörnsten
- abstract@[open-review](#): Most complex machine learning and modelling techniques are prone to over-fitting and may subsequently generalise poorly to future data. Artificial neural networks are no different in this regard and, despite having a level of implicit regularisation when trained with gradient descent, often require the aid of explicit regularisers. We introduce a new framework, Model Gradient Similarity (MGS), that (1) serves as a metric of regularisation, which can be used to monitor neural network training, (2) adds insight into how explicit regularisers, while derived from widely different principles, operate via the same mechanism underneath by increasing MGS, and (3) provides the basis for a new regularisation scheme which exhibits excellent performance, especially in challenging settings such as high levels of label noise or limited sample sizes.

## [Global Convergence of Direct Policy Search for State-Feedback \$\mathcal{H}^\infty\$ Robust Control: A Revisit of Nonsmooth Synthesis with Goldstein Subdifferential](#)

- Xingang Guo · Bin Hu
- abstract@[open-review](#): Direct policy search has been widely applied in modern reinforcement learning and continuous control. However, the performance of direct policy search on nonsmooth robust control synthesis has not been well understood. The optimal  $\mathcal{H}^\infty$  control framework aims at designing a policy to minimize the closed-loop  $\mathcal{H}^\infty$  norm, and is arguably the most important robust control paradigm. In this work, we show that direct policy search is guaranteed to find the global solution of the robust  $\mathcal{H}^\infty$  state-feedback control design problem. Notice that policy search for optimal  $\mathcal{H}^\infty$  control leads to a constrained nonconvex nonsmooth optimization problem where the nonconvex feasible set consists of all the policies stabilizing the closed-loop dynamics. We show that for this nonsmooth optimization problem, all Clarke stationary points are global minimum. Next, we identify the coercivity of the closed-loop  $\mathcal{H}^\infty$  objective function, and prove that the sublevel sets of the resultant policy search problem are compact. Based on these properties, we show that the Goldstein subdifferential method and its various implementable variants can be guaranteed to stay in the non-convex feasible set and eventually find the global optimal solution for the  $\mathcal{H}^\infty$  state-feedback synthesis problem. Our work builds a new connection between non-convex nonsmooth optimization theory and robust control, leading to the first global convergence result for direct policy search on optimal  $\mathcal{H}^\infty$  synthesis.

## [Differentially Private Graph Learning via Sensitivity-Bounded Personalized PageRank](#)

- Alessandro Epasto · Vahab Mirrokni · Bryan Perozzi · Anton Tsitsulin · Peilin Zhong
- abstract@[open-review](#): Personalized PageRank (PPR) is a fundamental tool in unsupervised learning of graph representations such as node ranking, labeling, and graph embedding. However, while data privacy is one of the most important recent concerns, existing PPR algorithms are not designed to protect user privacy. PPR is highly sensitive to the input graph edges: the difference of only one edge may cause a big change in the PPR vector, potentially leaking private user data. In this work, we propose an algorithm which outputs an approximate PPR and has provably bounded sensitivity to input edges. In addition, we prove that our algorithm achieves similar accuracy to non-private algorithms when the input graph has large degrees. Our sensitivity-bounded PPR directly implies private algorithms for several tools of graph learning, such as, differentially private (DP) PPR ranking, DP node classification, and DP node embedding. To complement our theoretical analysis, we also empirically verify the practical performances of our algorithms.

## [Sample Complexity of Learning Heuristic Functions for Greedy-Best-First and A\\* Search](#)

- Shinsaku Sakaue · Taihei Oki
- abstract@[open-review](#): Greedy best-first search (GBFS) and A search (A) are popular algorithms for path-finding on large graphs. Both use so-called heuristic functions, which estimate how close a vertex is to the goal. While heuristic functions have been handcrafted using domain knowledge, recent studies demonstrate that learning heuristic functions from data is effective in many applications. Motivated by this emerging approach, we study the sample complexity of learning heuristic functions for GBFS and A. We build on a recent framework called \textit{data-driven algorithm design} and evaluate the \textit{pseudo-dimension} of a class of utility functions that measure the performance of parameterized algorithms. Assuming that a vertex set of size  $n$  is fixed, we present  $\mathcal{O}(n \lg n)$  and  $\mathcal{O}(n^2 \lg n)$  upper bounds on the pseudo-dimensions for GBFS and A, respectively, parameterized by heuristic function values. The upper bound for A can be improved to  $\mathcal{O}(n^2 \lg d)$  if every vertex has a degree of at most  $d$  and to  $\mathcal{O}(n \lg n)$  if edge weights are integers bounded by  $\mathcal{O}(\text{poly}(n))$ . We also give  $\Omega(n)$  lower bounds for GBFS and A, which imply that our bounds for GBFS and A under the integer-weight condition are tight up to a  $\lg n$  factor. Finally, we discuss a case where the performance of A is measured by the suboptimality and show that we can sometimes obtain a better guarantee by combining a parameter-dependent worst-case bound with a sample complexity bound.

## [Statistical, Robustness, and Computational Guarantees for Sliced Wasserstein Distances](#)

- Sloan Nietert · Ziv Goldfeld · Ritwik Sadhu · Kengo Kato
- abstract@[open-review](#): Sliced Wasserstein distances preserve properties of classic Wasserstein distances while being more scalable for computation and estimation in high dimensions. The goal of this work is to quantify this scalability from three key aspects: (i) empirical convergence rates; (ii) robustness

to data contamination; and (iii) efficient computational methods. For empirical convergence, we derive fast rates with explicit dependence of constants on dimension, subject to log-concavity of the population distributions. For robustness, we characterize minimax optimal, dimension-free robust estimation rates, and show an equivalence between robust sliced 1-Wasserstein estimation and robust mean estimation. This enables lifting statistical and algorithmic guarantees available for the latter to the sliced 1-Wasserstein setting. Moving on to computational aspects, we analyze the Monte Carlo estimator for the average-sliced distance, demonstrating that larger dimension can result in faster convergence of the numerical integration error. For the max-sliced distance, we focus on a subgradient-based local optimization algorithm that is frequently used in practice, albeit without formal guarantees, and establish an  $\mathcal{O}(\epsilon^{-4})$  computational complexity bound for it. Our theory is validated with numerical experiments, which altogether provide a comprehensive quantitative account of the scalability question.

## [Increasing Confidence in Adversarial Robustness Evaluations](#)

- Roland S. Zimmermann · Wieland Brendel · Florian Tramer · Nicholas Carlini
- abstract@[open-review](#): Hundreds of defenses have been proposed to make deep neural networks robust against minimal (adversarial) input perturbations. However, only a handful of these defenses held up their claims because correctly evaluating robustness is extremely challenging: Weak attacks often fail to find adversarial examples even if they unknowingly exist, thereby making a vulnerable network look robust. In this paper, we propose a test to identify weak attacks, and thus weak defense evaluations. Our test slightly modifies a neural network to guarantee the existence of an adversarial example for every sample. Consequentially, any correct attack must succeed in breaking this modified network. For eleven out of thirteen previously-published defenses, the original evaluation of the defense fails our test, while stronger attacks that break these defenses pass it. We hope that attack unit tests - such as ours - will be a major component in future robustness evaluations and increase confidence in an empirical field that is currently riddled with skepticism.

## [RAMBO-RL: Robust Adversarial Model-Based Offline Reinforcement Learning](#)

- Marc Rigter · Bruno Lacerda · Nick Hawes
- abstract@[open-review](#): Offline reinforcement learning (RL) aims to find performant policies from logged data without further environment interaction. Model-based algorithms, which learn a model of the environment from the dataset and perform conservative policy optimisation within that model, have emerged as a promising approach to this problem. In this work, we present Robust Adversarial Model-Based Offline RL (RAMBO), a novel approach to model-based offline RL. To enforce conservatism, we formulate the problem as a two-player zero sum game against an adversarial environment model. The model is trained to minimise the value function while still accurately predicting the transitions in the dataset, forcing the policy to act conservatively in areas not covered by the dataset. To approximately solve the two-player game, we alternate between optimising the policy and adversarially optimising the model. The problem formulation that we address is theoretically grounded, resulting in a probably approximately correct (PAC) performance guarantee and a pessimistic value function which lower bounds the value function in the true environment. We evaluate our approach on widely studied offline RL benchmarks, and demonstrate that it outperforms existing state-of-the-art baselines.

## [Improving Generative Adversarial Networks via Adversarial Learning in Latent Space](#)

- Yang Li · Yichuan Mo · Liangliang Shi · Junchi Yan
- abstract@[open-review](#): For Generative Adversarial Networks which map a latent distribution to the target distribution, in this paper, we study how the sampling in latent space can affect the generation performance, especially for images. We observe that, as the neural generator is a continuous function, two close samples in latent space would be mapped into two nearby images, while their quality can differ much as the quality generally does not exhibit a continuous nature in pixel space. From such a continuous mapping function perspective, it is also possible that two distant latent samples can be mapped into two close images (if not exactly the same). In particular, if the latent samples are mapped in aggregation into a single mode, mode collapse occurs. Accordingly, we propose adding an implicit latent transform before the mapping function to improve latent  $z$  from its initial distribution, e.g., Gaussian. This is achieved using well-developed adversarial sample mining techniques, e.g. iterative fast gradient sign method (I-FGSM). We further propose new GAN training pipelines to obtain better generative mappings w.r.t quality and diversity by introducing targeted latent transforms into the bi-level optimization of GAN. Experimental results on visual data show that our method can effectively achieve improvement in both quality and diversity.

## [A sharp NMF result with applications in network modeling](#)

- Jiashun Jin
- abstract@[open-review](#): Given an  $n \times n$  non-negative rank- $K$  matrix  $\Omega$  where  $m$  eigenvalues are negative, when can we write  $\Omega = Z P Z'$  for non-negative matrices  $Z \in \mathbb{R}^{n \times K}$  and  $P \in \mathbb{R}^{K \times K}$ ? While most existing works focused on the case of  $m = 0$ , our primary interest is on the case of general  $m$ . With new proof ideas we develop, we present sharp results on when the NMF problem is solvable, which significantly extend existing results on this topic. The NMF problem is partially motivated by applications in network modeling. For a network with  $K$  communities, rank- $K$  models are popular, with many proposals. The DCMM model is a recent rank- $K$  model which is especially useful and interpretable in practice. To enjoy such properties, it is of interest to study when a rank- $K$  model can be rewritten as a DCMM model. Using our NMF results, we show that for a rank- $K$  model with parameters in the most interesting range, we can always rewrite it as a DCMM model.

## [Learning-Augmented Algorithms for Online Linear and Semidefinite Programming](#)

- Elena Grigorescu · Young-San Lin · Sandeep Silwal · Maoyuan Song · Samson Zhou
- abstract@[open-review](#): Semidefinite programming (SDP) is a unifying framework that generalizes both linear programming and quadratically-constrained quadratic programming, while also yielding efficient solvers, both in theory and in practice. However, there exist known impossibility results for approximating the optimal solution when constraints for covering SDPs arrive in an online fashion. In this paper, we study online linear and semidefinite covering programs in which the algorithm is augmented with advice from a possibly erroneous predictor. We show that if the predictor is accurate, we can efficiently bypass these impossibility results and achieve a constant-factor approximation to the optimal solution, i.e., consistency. On the other hand, if the predictor is inaccurate, under some technical conditions, we achieve results that match both the classical optimal upper bounds and the tight lower bounds up to constant factors, i.e., robustness. More broadly, we introduce a framework that extends both (1) the online set-cover problem augmented with machine-learning predictors, studied by Bamas, Maggiore, and Svensson (NeurIPS 2020), and (2) the online covering SDP problem, initiated by Elad, Kale, and Naor (ICALP 2016). Specifically, we obtain general online learning-augmented algorithms for covering linear programs with fractional advice and constraints, and initiate the study of learning-augmented algorithms for covering SDP problems. Our techniques are based on the primal-dual framework of Buchbinder and Naor (Mathematics of Operations Research, 34, 2009) and can be further adjusted to handle constraints where the variables lie in a bounded region, i.e., box constraints.

## [Bayesian Optimization over Discrete and Mixed Spaces via Probabilistic Reparameterization](#)

- Samuel Daulton · Xingchen Wan · David Eriksson · Maximilian Balandat · Eytan Bakshy · Michael A Osborne
- abstract@[open-review](#): Optimizing expensive-to-evaluate black-box functions of discrete (and potentially continuous) design parameters is a ubiquitous problem in scientific and engineering applications. Bayesian optimization (BO) is a popular sample-efficient method that selects promising designs to evaluate by optimizing an acquisition function (AF) over some domain with respect to a surrogate model. However, maximizing the AF over mixed or high-cardinality discrete search spaces is challenging as we cannot use standard gradient-based methods or evaluate the AF at every point in the search space. To address this issue, we propose using probabilistic reparameterization (PR). Instead of directly optimizing the AF over the search space containing discrete variables, we instead maximize the expectation of the AF over a probability distribution defined by continuous parameters. We prove

that under suitable proposal probability distributions, the BO policy that maximizes the probabilistic objective is the same as that which maximizes the AF, and therefore, PR enjoys the same regret bounds as the underlying AF. Moreover, our approach admits provably convergent global optimization of the AF (an often neglected requisite for commonly-used BO regret bounds) using scalable, unbiased estimators of both the probabilistic objective and its gradient. We validate our approach empirically and demonstrate state-of-the-art optimization performance on many real-world applications. Lastly, we showcase that PR is complementary to (and benefits) recent work and naturally generalizes to settings with multiple objectives and black-box constraints.

## [Few-Shot Parameter-Efficient Fine-Tuning is Better and Cheaper than In-Context Learning](#)

- Haokun Liu · Derek Tam · Mohammed Muqeeth · Jay Mohta · Tenghao Huang · Mohit Bansal · Colin Raffel
- abstract@[open-review](#): Few-shot in-context learning (ICL) enables pre-trained language models to perform a previously-unseen task without any gradient-based training by feeding a small number of training examples as part of the input. ICL incurs substantial computational, memory, and storage costs because it involves processing all of the training examples every time a prediction is made. Parameter-efficient fine-tuning (PEFT) (e.g. adapter modules, prompt tuning, sparse update methods, etc.) offers an alternative paradigm where a small set of parameters are trained to enable a model to perform the new task. In this paper, we rigorously compare few-shot ICL and PEFT and demonstrate that the latter offers better accuracy as well as dramatically lower computational costs. Along the way, we introduce a new PEFT method called (IA)<sup>3</sup> that scales activations by learned vectors, attaining stronger performance while only introducing a relatively tiny amount of new parameters. We also propose a simple recipe based on the T0 model called T-Few that can be applied to new tasks without task-specific tuning or modifications. We validate the effectiveness of T-Few on completely unseen tasks by applying it to the RAFT benchmark, attaining super-human performance for the first time and outperforming the state-of-the-art by 6% absolute. All of the code used in our experiments will be publicly available.

## [Rethinking Individual Global Max in Cooperative Multi-Agent Reinforcement Learning](#)

- Yaochu Jin · Yitian Hong · Yang Tang
- abstract@[open-review](#): In cooperative multi-agent reinforcement learning, centralized training and decentralized execution (CTDE) has achieved remarkable success. Individual Global Max (IGM) decomposition, which is an important element of CTDE, measures the consistency between local and joint policies. The majority of IGM-based research focuses on how to establish this consistent relationship, but little attention has been paid to examining IGM's potential flaws. In this work, we reveal that the IGM condition is a lossy decomposition, and the error of lossy decomposition will accumulate in hypernetwork-based methods. To address the above issue, we propose to adopt an imitation learning strategy to separate the lossy decomposition from Bellman iterations, thereby avoiding error accumulation. The proposed strategy is theoretically proved and empirically verified on the StarCraft Multi-Agent Challenge benchmark problem with zero sight view. The results also confirm that the proposed method outperforms state-of-the-art IGM-based approaches.

## [Reduction Algorithms for Persistence Diagrams of Networks: CoralTDA and PrunIT](#)

- Cuneyt G Akcora · Murat Kantacioglu · Yulia Gel · Baris Coskunuzer
- abstract@[open-review](#): Topological data analysis (TDA) delivers invaluable and complementary information on the intrinsic properties of data inaccessible with conventional methods. However, high computational costs remain the primary roadblock hindering the successful application of TDA in real-world studies, particularly in conjunction with machine learning on large complex networks. Indeed, most modern networks such as citation, blockchain, and online social networks often have hundreds of thousands of vertices, making the application of existing TDA methods infeasible. We develop two new, remarkably simple but effective algorithms to compute the exact persistence diagrams of large graphs to address this major TDA limitation. First, we prove that  $(k+1)$ -core of a graph  $G$  is sufficient to compute its  $k^{\{th\}}$  persistence diagram,  $\text{PD}_k(\text{CG})$ . Second, we introduce a pruning algorithm for graphs to compute their persistence diagrams by removing the dominated vertices. Our experiments on large networks indicate that the new approach can achieve computational gains up to 95%. The developed framework provides the first bridge between the graph theory and TDA, with applications in machine learning of large complex networks.

## [NaturalProver: Grounded Mathematical Proof Generation with Language Models](#)

- Sean Welleck · Jiacheng Liu · Ximing Lu · Hannaneh Hajishirzi · Yejin Choi
- abstract@[open-review](#): Theorem proving in natural mathematical language –“the mixture of symbolic and natural language used by humans” plays a central role in mathematical advances and education, and tests aspects of reasoning that are core to intelligence. Yet it has remained underexplored with modern generative models. We study large-scale language models on two new generation tasks: suggesting the next step in a mathematical proof, and full proof generation. We develop NaturalProver, a language model that generates proofs by conditioning on background references (e.g. theorems and definitions that are either retrieved or human-provided), and optionally enforces their presence with constrained decoding. On theorems from the NaturalProofs benchmark, NaturalProver improves the quality of next-step suggestions and generated proofs over fine-tuned GPT-3, according to human evaluations from university-level mathematics students. NaturalProver is capable of proving some theorems that require short (2-6 step) proofs, and providing next-step suggestions that are rated as correct and useful over 40% of the time, which is to our knowledge the first demonstration of these capabilities using neural language models.

## [CLOOB: Modern Hopfield Networks with InfoLOOB Outperform CLIP](#)

- Andreas Färst · Elisabeth Rumetschöfer · Johannes Lehner · Viet Tran · Fei Tang · Hubert Ramsauer · David Kreil · Michael Kopp · Gápter Klambauer · Angela Bitto · Sepp Hochreiter
- abstract@[open-review](#): CLIP yielded impressive results on zero-shot transfer learning tasks and is considered as a foundation model like BERT or GPT3. CLIP vision models that have a rich representation are pre-trained using the InfoNCE objective and natural language supervision before they are fine-tuned on particular tasks. Though CLIP excels at zero-shot transfer learning, it suffers from an explaining away problem, that is, it focuses on one or few features, while neglecting other relevant features. This problem is caused by insufficiently extracting the covariance structure in the original multi-modal data. We suggest to use modern Hopfield networks to tackle the problem of explaining away. Their retrieved embeddings have an enriched covariance structure derived from co-occurrences of features in the stored embeddings. However, modern Hopfield networks increase the saturation effect of the InfoNCE objective which hampers learning. We propose to use the InfoLOOB objective to mitigate this saturation effect. We introduce the novel “Contrastive Leave One Out Boost” (CLOOB), which uses modern Hopfield networks for covariance enrichment together with the InfoLOOB objective. In experiments we compare CLOOB to CLIP after pre-training on the Conceptual Captions and the YFCC dataset with respect to their zero-shot transfer learning performance on other datasets. CLOOB consistently outperforms CLIP at zero-shot transfer learning across all considered architectures and datasets.

## [Friendly Noise against Adversarial Noise: A Powerful Defense against Data Poisoning Attack](#)

- Tian Yu Liu · Yu Yang · Baharan Mirzasoleiman
- abstract@[open-review](#): Data poisoning attacks modify a subset of training examples by small adversarial perturbations to change the prediction of certain test-time data. Existing defense mechanisms are not desirable to deploy in practice, as they often drastically harm the generalization performance, or are attack-specific and prohibitively slow to apply. Here, we propose a simple but highly effective approach that unlike existing methods breaks various types of poisoning attacks with the slightest drop in the generalization performance. We make the key observation that attacks exploit sharp loss regions to craft adversarial perturbations which can substantially alter examples' gradient or representations under small perturbations. To break poisoning attacks, our

approach comprises two components: an optimized friendly noise that is generated to maximally perturb examples without degrading the performance, and a random varying noise component. The first component takes examples farther away from the sharp loss regions, and the second component smooths out the loss landscape. The combination of both components builds a very light-weight but extremely effective defense against the most powerful triggerless and backdoor poisoning attacks, including Gradient Matching, Bulls-eye Polytope, and Sleeper Agent. We show that our friendly noise is transferable to other architectures, and adaptive attacks cannot break our defense due to its random noise component.

## [An \$\alpha\$ -No-Regret Algorithm For Graphical Bilinear Bandits](#)

- Geovani Rizk · Igor Colin · Albert Thomas · Rida Laraki · Yann Chevaleyre
- abstract@[open-review](#): We propose the first regret-based approach to the \emph{Graphical Bilinear Bandits} problem, where \$n\$ agents in a graph play a stochastic bilinear bandit game with each of their neighbors. This setting reveals a combinatorial NP-hard problem that prevents the use of any existing regret-based algorithm in the (bi-)linear bandit literature. In this paper, we fill this gap and present the first regret-based algorithm for graphical bilinear bandits using the principle of optimism in the face of uncertainty. Theoretical analysis of this new method yields an upper bound of  $\tilde{O}(\sqrt{T})$  on the  $\alpha$ -regret and evidences the impact of the graph structure on the rate of convergence. Finally, we show through various experiments the validity of our approach.

## [EAGER: Asking and Answering Questions for Automatic Reward Shaping in Language-guided RL](#)

- Thomas Carta · Pierre-Yves Oudeyer · Olivier Sigaud · Sylvain Lamprier
- abstract@[open-review](#): Reinforcement learning (RL) in long horizon and sparse reward tasks is notoriously difficult and requires a lot of training steps. A standard solution to speed up the process is to leverage additional reward signals, shaping it to better guide the learning process. In the context of language-conditioned RL, the abstraction and generalisation properties of the language input provide opportunities for more efficient ways of shaping the reward. In this paper, we leverage this idea and propose an automated reward shaping method where the agent extracts auxiliary objectives from the general language goal. These auxiliary objectives use a question generation (QG) and a question answering (QA) system: they consist of questions leading the agent to try to reconstruct partial information about the global goal using its own trajectory. When it succeeds, it receives an intrinsic reward proportional to its confidence in its answer. This incentivizes the agent to generate trajectories which unambiguously explain various aspects of the general language goal. Our experimental study using various BabyAI environments shows that this approach, which does not require engineer intervention to design the auxiliary objectives, improves sample efficiency by effectively directing the exploration.

## [Adv-Attribute: Inconspicuous and Transferable Adversarial Attack on Face Recognition](#)

- Shuai Jia · Bangjie Yin · Taiping Yao · Shouhong Ding · Chunhua Shen · Xiaokang Yang · Chao Ma
- abstract@[open-review](#): Deep learning models have shown their vulnerability when dealing with adversarial attacks. Existing attacks almost perform on low-level instances, such as pixels and super-pixels, and rarely exploit semantic clues. For face recognition attacks, existing methods typically generate the  $l_p$ -norm perturbations on pixels, however, resulting in low attack transferability and high vulnerability to denoising defense models. In this work, instead of performing perturbations on the low-level pixels, we propose to generate attacks through perturbing on the high-level semantics to improve attack transferability. Specifically, a unified flexible framework, Adversarial Attributes (Adv-Attribute), is designed to generate inconspicuous and transferable attacks on face recognition, which crafts the adversarial noise and adds it into different attributes based on the guidance of the difference in face recognition features from the target. Moreover, the importance-aware attribute selection and the multi-objective optimization strategy are introduced to further ensure the balance of stealthiness and attacking strength. Extensive experiments on the FFHQ and CelebA-HQ datasets show that the proposed Adv-Attribute method achieves the state-of-the-art attacking success rates while maintaining better visual effects against recent attack methods.

## [RankFeat: Rank-1 Feature Removal for Out-of-distribution Detection](#)

- Yue Song · Nicu Sebe · Wei Wang
- abstract@[open-review](#): The task of out-of-distribution (OOD) detection is crucial for deploying machine learning models in real-world settings. In this paper, we observe that the singular value distributions of the in-distribution (ID) and OOD features are quite different: the OOD feature matrix tends to have a larger dominant singular value than the ID feature, and the class predictions of OOD samples are largely determined by it. This observation motivates us to propose RankFeat, a simple yet effective post hoc approach for OOD detection by removing the rank-1 matrix composed of the largest singular value and the associated singular vectors from the high-level feature. RankFeat achieves state-of-the-art performance and reduces the average false positive rate (FPR95) by 17.90% compared with the previous best method. Extensive ablation studies and comprehensive theoretical analyses are presented to support the empirical results.

## [Global convergence of ResNets: From finite to infinite width using linear parameterization](#)

- Raphaël Barboni · Gabriel Peyré · François-Xavier Vialard
- abstract@[open-review](#): Overparameterization is a key factor in the absence of convexity to explain global convergence of gradient descent (GD) for neural networks. Beside the well studied lazy regime, infinite width (mean field) analysis has been developed for shallow networks, using on convex optimization technics. To bridge the gap between the lazy and mean field regimes, we study Residual Networks (ResNets) in which the residual block has linear parameterization while still being nonlinear. Such ResNets admit both infinite depth and width limits, encoding residual blocks in a Reproducing Kernel Hilbert Space (RKHS). In this limit, we prove a local Polyak-Lojasiewicz inequality. Thus, every critical point is a global minimizer and a local convergence result of GD holds, retrieving the lazy regime. In contrast with other mean-field studies, it applies to both parametric and non-parametric cases under an expressivity condition on the residuals. Our analysis leads to a practical and quantified recipe: starting from a universal RKHS, Random Fourier Features are applied to obtain a finite dimensional parameterization satisfying with high-probability our expressivity condition.

## [Decoupling Features in Hierarchical Propagation for Video Object Segmentation](#)

- Zongxin Yang · Yi Yang
- abstract@[open-review](#): This paper focuses on developing a more effective method of hierarchical propagation for semi-supervised Video Object Segmentation (VOS). Based on vision transformers, the recently-developed AOT approach introduces hierarchical propagation into VOS and has shown promising results. The hierarchical propagation can gradually propagate information from past frames to the current frame and transfer the current frame feature from object-agnostic to object-specific. However, the increase of object-specific information will inevitably lead to the loss of object-agnostic visual information in deep propagation layers. To solve such a problem and further facilitate the learning of visual embeddings, this paper proposes a Decoupling Features in Hierarchical Propagation (DeAOT) approach. Firstly, DeAOT decouples the hierarchical propagation of object-agnostic and object-specific embeddings by handling them in two independent branches. Secondly, to compensate for the additional computation from dual-branch propagation, we propose an efficient module for constructing hierarchical propagation, i.e., Gated Propagation Module, which is carefully designed with single-head attention. Extensive experiments show that DeAOT significantly outperforms AOT in both accuracy and efficiency. On YouTube-VOS, DeAOT can achieve 86.0% at 22.4fps and 82.0% at 53.4fps. Without test-time augmentations, we achieve new state-of-the-art performance on four benchmarks, i.e., YouTube-VOS (86.2%), DAVIS 2017 (86.2%), DAVIS 2016 (92.9%), and VOT 2020 (0.622 EAO). The code will be publicly available.

## [PKD: General Distillation Framework for Object Detectors via Pearson Correlation Coefficient](#)

- Weihan Cao · Jianfei Gao · Anda Cheng · Ke Cheng · Yifan Zhang · Jian Cheng
- abstract@[open-review](#): Knowledge distillation(KD) is a widely-used technique to train compact models in object detection. However, there is still a lack of study on how to distill between heterogeneous detectors. In this paper, we empirically find that better FPN features from a heterogeneous teacher detector can help the student although their detection heads and label assignments are different. However, directly aligning the feature maps to distill detectors suffers from two problems. First, the difference in feature magnitude between the teacher and the student could enforce overly strict constraints on the student. Second, the FPN stages and channels with large feature magnitude from the teacher model could dominate the gradient of distillation loss, which will overwhelm the effects of other features in KD and introduce much noise. To address the above issues, we propose to imitate features with Pearson Correlation Coefficient to focus on the relational information from the teacher and relax constraints on the magnitude of the features. Our method consistently outperforms the existing detection KD methods and works for both homogeneous and heterogeneous student-teacher pairs. Furthermore, it converges faster. With a powerful MaskRCNN-Swin detector as the teacher, ResNet-50 based RetinaNet and FCOS achieve 41.5% and 43.9% \$mAP\$ on COCO2017, which are 4.1% and 4.8% higher than the baseline, respectively.

## [CHIMLE: Conditional Hierarchical IMLE](#)

- Shichong Peng · Seyed Alireza Moazenipourasil · Ke Li
- abstract@[open-review](#): A persistent challenge in conditional image synthesis has been generating diverse output images from the same input image due to the problem of mode collapse. Implicit Maximum Likelihood Estimation (IMLE) is a recently proposed alternative that aims to address this issue. IMLE uses the same generator as GANs but adopts a different objective function which ensures each observed image has a generated sample nearby. To generate high-fidelity images, prior IMLE-based methods require a large number of samples. Doing so is expensive, and so this limits image fidelity in practice. In this paper, we propose a new method to get around this limitation, which we dub Conditional Hierarchical IMLE (CHIMLE), which can generate high-fidelity images without requiring many samples. We show on multiple tasks that CHIMLE significantly improves generated image fidelity, as demonstrated by a reduction in FrÃ©chet Inception Distance (FID) by 36.9% on average compared to the prior best IMLE-based method.

## [Moment Distributionally Robust Tree Structured Prediction](#)

- Yesu Li · Danyal Saeed · Xinhua Zhang · Brian Ziebart · Kevin Gimpel
- abstract@[open-review](#): Structured prediction of tree-shaped objects is heavily studied under the name of syntactic dependency parsing. Current practice based on maximum likelihood or margin is either agnostic to or inconsistent with the evaluation loss. Risk minimization alleviates the discrepancy between training and test objectives but typically induces a non-convex problem. These approaches adopt explicit regularization to combat overfitting without probabilistic interpretation. We propose a moment-based distributionally robust optimization approach for tree structured prediction, where the worst-case expected loss over a set of distributions within bounded moment divergence from the empirical distribution is minimized. We develop efficient algorithms for arborescences and other variants of trees. We derive Fisher consistency, convergence rates and generalization bounds for the proposed method. We evaluate empirical effectiveness on dependency parsing benchmarks.

## [Blessing of Nonconvexity in Deep Linear Models: Depth Flattens the Optimization Landscape Around the True Solution](#)

- Jianhao Ma · Salar Fattah
- abstract@[open-review](#): This work characterizes the effect of depth on the optimization landscape of linear regression, showing that, despite their nonconvexity, deeper models have more desirable optimization landscapes. We consider a robust and over-parameterized setting, where a subset of measurements are grossly corrupted with noise, and the true linear model is captured via an  $N$ -layer linear neural network. On the negative side, we show that this problem does not have a benign landscape: given any  $N \geq 1$ , with constant probability, there exists a solution corresponding to the ground truth that is neither local nor global minimum. However, on the positive side, we prove that, for any  $N$ -layer model with  $N \geq 2$ , a simple sub-gradient method becomes oblivious to such “problematic” solutions; instead, it converges to a balanced solution that is not only close to the ground truth but also enjoys a flat local landscape, thereby eschewing the need for “early stopping”. Lastly, we empirically verify that the desirable optimization landscape of deeper models extends to other robust learning tasks, including deep matrix recovery and deep ReLU networks with  $\ell_1$ -loss.

## [KSD Aggregated Goodness-of-fit Test](#)

- Antonin Schrab · Benjamin Guedj · Arthur Gretton
- abstract@[open-review](#): We investigate properties of goodness-of-fit tests based on the Kernel Stein Discrepancy (KSD). We introduce a strategy to construct a test, called KSDAgg, which aggregates multiple tests with different kernels. KSDAgg avoids splitting the data to perform kernel selection (which leads to a loss in test power), and rather maximises the test power over a collection of kernels. We provide theoretical guarantees on the power of KSDAgg: we show it achieves the smallest uniform separation rate of the collection, up to a logarithmic term. KSDAgg can be computed exactly in practice as it relies either on a parametric bootstrap or on a wild bootstrap to estimate the quantiles and the level corrections. In particular, for the crucial choice of bandwidth of a fixed kernel, it avoids resorting to arbitrary heuristics (such as median or standard deviation) or to data splitting. We find on both synthetic and real-world data that KSDAgg outperforms other state-of-the-art adaptive KSD-based goodness-of-fit testing procedures.

## [GFlowCausal: Generative Flow Networks for Causal Discovery](#)

- Wenqian Li · Yinchuan Li · Shengyu Zhu · Shao Yunfeng · Jianye Hao · Yan Pang
- abstract@[open-review](#): Causal discovery aims to uncover causal relationships among a set of variables. Score-based approaches mainly focus on searching for the best Directed Acyclic Graph (DAG) based on a predefined score function. However, most of them are not applicable on a large scale due to the limited searchability. Inspired by the active learning in generative flow networks, we propose a novel approach to learn a DAG from observational data, which called GFlowDAG. It converts the graph search problem to a generation problem, in which direct edges are added gradually. GFlowDAG aims to learn the best policy to generate high-reward DAGs by sequential actions with probabilities proportional to predefined rewards. We propose a plug-and-play module based on transitive closure to ensure efficiently sampling. Theoretical analysis shows that this module could guarantee acyclicity properties effectively and the consistency between final states and fully-connected graphs. We conduct extensive experiments on both synthetic and real datasets, and results show the proposed approach to be superior and also performs well in a large-scale setting.

## [Towards Theoretically Inspired Neural Initialization Optimization](#)

- Yibo Yang · Hong Wang · Haobo Yuan · Zhouchen Lin
- abstract@[open-review](#): Automated machine learning has been widely explored to reduce human efforts in designing neural architectures and looking for proper hyperparameters. In the domain of neural initialization, however, similar automated techniques have rarely been studied. Most existing initialization methods are handcrafted and highly dependent on specific architectures. In this paper, we propose a differentiable quantity, named GradCosine, with theoretical insights to evaluate the initial state of a neural network. Specifically, GradCosine is the cosine similarity of sample-wise gradients with respect to the initialized parameters. By analyzing the sample-wise optimization landscape, we show that both the training and test performance of a network can be improved by maximizing GradCosine under gradient norm constraint. Based on this observation, we further propose the neural initialization optimization (NIO) algorithm. Generalized from the sample-wise analysis into the real batch setting, NIO is able to automatically look for a better initialization with negligible cost compared with the training time. With NIO, we improve the classification performance of a variety of neural

architectures on CIFAR10, CIFAR-100, and ImageNet. Moreover, we find that our method can even help to train large vision Transformer architecture without warmup.

## [To update or not to update? Neurons at equilibrium in deep models](#)

- Andrea Bragagnolo · Enzo Tartaglione · Marco Grangetto
- abstract@[open-review](#): Recent advances in deep learning optimization showed that, with some a-posteriori information on fully-trained models, it is possible to match the same performance by simply training a subset of their parameters. Such a discovery has a broad impact from theory to applications, driving the research towards methods to identify the minimum subset of parameters to train without look-ahead information exploitation. However, the methods proposed do not match the state-of-the-art performance, and rely on unstructured sparsely connected models. In this work we shift our focus from the single parameters to the behavior of the whole neuron, exploiting the concept of neuronal equilibrium (NEq). When a neuron is in a configuration at equilibrium (meaning that it has learned a specific input-output relationship), we can halt its update; on the contrary, when a neuron is at non-equilibrium, we let its state evolve towards an equilibrium state, updating its parameters. The proposed approach has been tested on different state-of-the-art learning strategies and tasks, validating NEq and observing that the neuronal equilibrium depends on the specific learning setup.

## [Reducing Confidence Along Adversarial Directions: Maximizing Entropy on Self-Generated Perturbations](#)

- Amirth Setlur · Benjamin Eysenbach · Virginia Smith · Sergey Levine
- abstract@[open-review](#): Supervised learning methods trained with maximum likelihood objectives often overfit on training data. Most regularizers that prevent overfitting look to increase confidence on additional examples (e.g., data augmentation, adversarial training), or reduce it on training data (e.g., label smoothing). In this work we propose a complementary regularization strategy that reduces confidence on self-generated examples. We call our regularizer RCAD: Reducing Confidence along Adversarial Directions, and the key idea behind it is to reduce confidence on out-of-distribution examples lying along directions adversarially chosen to increase training loss. In contrast to adversarial training, our method does not try to robustify the model to output the original label, but rather regularizes it to have reduced confidence on points generated using much larger perturbations than in conventional adversarial training. RCAD can be easily integrated into training pipelines with a few lines of code. Despite its simplicity, we find on many classification benchmarks that RCAD can be added to existing techniques (e.g., label smoothing, MixUp training) to increase test accuracy by 1-3% in absolute value, with more significant gains in the low data regime. We also provide a theoretical analysis that helps to explain these benefits in simplified settings, showing that RCAD can provably help the model unlearn spurious features in the training data.

## [Self-supervised learning of brain dynamics from broad neuroimaging data](#)

- Armin Thomas · Christopher RÃ© · Russell Poldrack
- abstract@[open-review](#): Self-supervised learning techniques are celebrating immense success in natural language processing (NLP) by enabling models to learn from broad language data at unprecedented scales. Here, we aim to leverage the success of these techniques for mental state decoding, where researchers aim to identify specific mental states (e.g., the experience of anger or joy) from brain activity. To this end, we devise a set of novel self-supervised learning frameworks for neuroimaging data based on prominent learning frameworks in NLP. At their core, these frameworks learn the dynamics of brain activity by modeling sequences of activity akin to how NLP models sequences of text. We evaluate the frameworks by pre-training models on a broad neuroimaging dataset spanning functional Magnetic Resonance Imaging data from 11,980 experimental runs of 1,726 individuals across 34 datasets and subsequently adapting the pre-trained models to two benchmark mental state decoding datasets. The pre-trained models transfer well, generally outperforming baseline models trained from scratch, while models trained in a learning framework based on causal language modeling clearly outperform the others.

## [Stochastic Halpern Iteration with Variance Reduction for Stochastic Monotone Inclusions](#)

- Xufeng Cai · Chaobing Song · CristÃ³bal GuzmÃ¡n · Jelena Diakonikolas
- abstract@[open-review](#): We study stochastic monotone inclusion problems, which widely appear in machine learning applications, including robust regression and adversarial learning. We propose novel variants of stochastic Halpern iteration with recursive variance reduction. In the cocoercive---and more generally Lipschitz-monotone---setup, our algorithm attains  $\|\epsilon\|$  norm of the operator with  $\mathcal{O}(\frac{1}{\epsilon^3})$  stochastic operator evaluations, which significantly improves over state of the art  $\mathcal{O}(\frac{1}{\epsilon^4})$  stochastic operator evaluations required for existing monotone inclusion solvers applied to the same problem classes. We further show how to couple one of the proposed variants of stochastic Halpern iteration with a scheduled restart scheme to solve stochastic monotone inclusion problems with  $\mathcal{O}(\frac{\log(1/\epsilon)}{\epsilon^2})$  stochastic operator evaluations under additional sharpness or strong monotonicity assumptions. Finally, we argue via reductions between different problem classes that our stochastic oracle complexity bounds are tight up to logarithmic factors in terms of their  $\|\epsilon\|$ -dependence.

## [Coordinate Linear Variance Reduction for Generalized Linear Programming](#)

- Chaobing Song · Cheuk Yin Lin · Stephen Wright · Jelena Diakonikolas
- abstract@[open-review](#): We study a class of generalized linear programs (GLP) in a large-scale setting, which includes simple, possibly nonsmooth convex regularizer and simple convex set constraints. By reformulating (GLP) as an equivalent convex-concave min-max problem, we show that the linear structure in the problem can be used to design an efficient, scalable first-order algorithm, to which we give the name Coordinate Linear Variance Reduction (CLVR; pronounced ``clever''). CLVR yields improved complexity results for (GLP) that depend on the max row norm of the linear constraint matrix in (GLP) rather than the spectral norm. When the regularization terms and constraints are separable, CLVR admits an efficient lazy update strategy that makes its complexity bounds scale with the number of nonzero elements of the linear constraint matrix in (GLP) rather than the matrix dimensions. On the other hand, for the special case of linear programs, by exploiting sharpness, we propose a restart scheme for CLVR to obtain empirical linear convergence. Then we show that Distributionally Robust Optimization (DRO) problems with ambiguity sets based on both  $f$ -divergence and Wasserstein metrics can be reformulated as (GLPs) by introducing sparsely connected auxiliary variables. We complement our theoretical guarantees with numerical experiments that verify our algorithm's practical effectiveness, in terms of wall-clock time and number of data passes.

## [Adversarial training for high-stakes reliability](#)

- Daniel Ziegler · Seraphina Nix · Lawrence Chan · Tim Bauman · Peter Schmidt-Nielsen · Tao Lin · Adam Scherlis · Noa Nabeshima · Benjamin Weinstein-Raun · Daniel de Haas · Buck Shlegeris · Nate Thomas
- abstract@[open-review](#): In the future, powerful AI systems may be deployed in high-stakes settings, where a single failure could be catastrophic. One technique for improving AI safety in high-stakes settings is adversarial training, which uses an adversary to generate examples to train on in order to achieve better worst-case performance. In this work, we used a safe language generation task ("`avoid injuries") as a testbed for achieving high reliability through adversarial training. We created a series of adversarial training techniques---including a tool that assists human adversaries---to find and eliminate failures in a classifier that filters text completions suggested by a generator. In our task, we determined that we can set very conservative classifier thresholds without significantly impacting the quality of the filtered outputs. We found that adversarial training significantly increased robustness to the adversarial attacks that we trained on--- tripling the time to find adversarial examples without tools and doubling the time with our tool (from 13 to 26 minutes)---without affecting in-distribution performance. We hope to see further work in the high-stakes reliability setting, including more powerful tools for enhancing human adversaries and better ways to measure high levels of reliability, until we can confidently rule out the possibility of catastrophic deployment-time failures of powerful models.

## [How Powerful are K-hop Message Passing Graph Neural Networks](#)

- Jiarui Feng · Yixin Chen · Fuhai Li · Anindya Sarkar · Muhan Zhang
- abstract@[open-review](#): The most popular design paradigm for Graph Neural Networks (GNNs) is 1-hop message passing---aggregating features from 1-hop neighbors repeatedly. However, the expressive power of 1-hop message passing is bounded by the Weisfeiler-Lehman (1-WL) test. Recently, researchers extended 1-hop message passing to  $K$ -hop message passing by aggregating information from  $K$ -hop neighbors of nodes simultaneously. However, there is no work on analyzing the expressive power of  $K$ -hop message passing. In this work, we theoretically characterize the expressive power of  $K$ -hop message passing. Specifically, we first formally differentiate two kinds of kernels of  $K$ -hop message passing which are often misused in previous works. We then characterize the expressive power of  $K$ -hop message passing by showing that it is more powerful than 1-hop message passing. Despite the higher expressive power, we show that  $K$ -hop message passing still cannot distinguish some simple regular graphs. To further enhance its expressive power, we introduce a KP-GNN framework, which improves  $K$ -hop message passing by leveraging the peripheral subgraph information in each hop. We prove that KP-GNN can distinguish almost all regular graphs including some distance regular graphs which could not be distinguished by previous distance encoding methods. Experimental results verify the expressive power and effectiveness of KP-GNN. KP-GNN achieves competitive results across all benchmark datasets.

## [K-LITE: Learning Transferable Visual Models with External Knowledge](#)

- Sheng Shen · Chunyuan Li · Xiaowei Hu · Yujia Xie · Jianwei Yang · Pengchuan Zhang · Zhe Gan · Lijuan Wang · Lu Yuan · Ce Liu · Kurt Keutzer · Trevor Darrell · Anna Rohrbach · Jianfeng Gao
- abstract@[open-review](#): The new generation of state-of-the-art computer vision systems are trained from natural language supervision, ranging from simple object category names to descriptive captions. This form of supervision ensures high generality and usability of the learned visual models, based on the broad concept coverage achieved through large-scale data collection process. Alternatively, we argue that learning with external knowledge about images is a promising way which leverages a much more structured source of supervision and offers sample efficiency. In this paper, we propose K-LITE (Knowledge-augmented Language-Image Training and Evaluation), a simple strategy to leverage external knowledge for building transferable visual systems: In training, it enriches entities in natural language with WordNet and Wiktionary knowledge, leading to an efficient and scalable approach to learning image representations that uses knowledge about the visual concepts; In evaluation, the natural language is also augmented with external knowledge and then used to reference learned visual concepts (or describe new ones) to enable zero-shot and few-shot transfer of the pre-trained models. We study the performance of K-LITE on two important computer vision problems, image classification and object detection, benchmarking on 20 and 13 different existing datasets, respectively. The proposed knowledge-augmented models show significant improvement in transfer learning performance over existing methods

## [Statistical Learning and Inverse Problems: An Stochastic Gradient Approach](#)

- Yuri Fonseca
- abstract@[open-review](#): Inverse problems are paramount in Science and Engineering. In this paper, we consider the setup of Statistical Inverse Problem (SIP) and demonstrate how Stochastic Gradient Descent (SGD) algorithms can be used in the linear SIP setting. We provide consistency and finite sample bounds for the excess risk. We also propose a modification for the SGD algorithm where we leverage machine learning methods to smooth the stochastic gradients and improve empirical performance. We exemplify the algorithm in a setting of great interest nowadays: the Functional Linear Regression model. In this case we consider a synthetic data example and examples with a real data classification problem.

## [Network change point localisation under local differential privacy](#)

- Mengchu Li · Tom Berrett · Yi Yu
- abstract@[open-review](#): Network data are ubiquitous in our daily life, containing rich but often sensitive information. In this paper, we expand the current static analysis of privatised networks to a dynamic framework by considering a sequence of networks with potential change points. We investigate the fundamental limits in consistently localising change points under both node and edge privacy constraints, demonstrating interesting phase transition in terms of the signal-to-noise ratio condition, accompanied by polynomial-time algorithms. The private signal-to-noise ratio conditions quantify the costs of the privacy for change point localisation problems and exhibit a different scaling in the sparsity parameter compared to the non-private counterparts. Our algorithms are shown to be optimal under the edge LDP constraint up to log factors. Under node LDP constraint, a gap exists between our upper bound and lower bound and we leave it as an interesting open problem, echoing the challenges in high-dimensional statistical inference under LDP constraints.

## [Efficient Submodular Optimization under Noise: Local Search is Robust](#)

- Lingxiao Huang · Yuyi Wang · Chunxue Yang · Huanjian Zhou
- abstract@[open-review](#): The problem of monotone submodular maximization has been studied extensively due to its wide range of applications. However, there are cases where one can only access the objective function in a distorted or noisy form because of the uncertain nature or the errors involved in the evaluation. This paper considers the problem of constrained monotone submodular maximization with noisy oracles introduced by Hassidim and Singer (2017). For a cardinality constraint, we propose an algorithm achieving a near-optimal  $(1-1/e-O(\epsilon))$ -approximation guarantee (for arbitrary  $\epsilon > 0$ ) with only a polynomial number of queries to the noisy value oracle, which improves the exponential query complexity of Singer and Hassidim (2018). For general matroid constraints, we show the first constant approximation algorithm in the presence of noise. Our main approaches are to design a novel local search framework that can handle the effect of noise and to construct certain smoothing surrogate functions for noise reduction.

## [Reconciling intrinsic rewards via constrained policy optimization](#)

- Eric Chen · Zhang-Wei Hong · Joni Pajarinen · Pulkit Agrawal
- abstract@[open-review](#): State-of-the-art reinforcement learning (RL) algorithms typically use random sampling (e.g.,  $\epsilon$ -greedy) for exploration, but this method fails in hard exploration tasks like Montezuma's Revenge. To address the challenge of exploration, prior works incentivize the agent to visit novel states using an exploration bonus (also called intrinsic rewards), which led to excellent results on some hard exploration tasks. However, recent studies show that on many other tasks intrinsic rewards can bias policy optimization leading to poor performance compared to optimizing only the environment reward. The low-performance results from the agent seeking intrinsic rewards and performing unnecessary exploration even when sufficient environment reward is provided. This inconsistency in performance across tasks prevents widespread use of intrinsic rewards with RL algorithms. We propose a principled constrained policy optimization procedure to eliminate the detrimental effects of intrinsic rewards while preserving their merits when applicable. Our method automatically tunes the importance of intrinsic reward: it suppresses intrinsic rewards when they are not needed and increases them when exploration is required. The end result is a superior exploration algorithm that does not require manual tuning to balance intrinsic rewards against environment rewards. Experimental results across 61 Atari games validate our claim.

## [Bridge the Gap Between Architecture Spaces via A Cross-Domain Predictor](#)

- Yuqiao Liu · Yehui Tang · Zeqiong Lv · Yunhe Wang · Yanan Sun
- abstract@[open-review](#): Neural Architecture Search (NAS) can automatically design promising neural architectures without artificial experience. Though it achieves great success, prohibitively high search cost is required to find a high-performance architecture, which blocks its practical implementation.

Neural predictor can directly evaluate the performance of neural networks based on their architectures and thereby save much budget. However, existing neural predictors require substantial annotated architectures trained from scratch, which still consume many computational resources. To solve this issue, we propose a Cross-Domain Predictor (CDP), which is trained based on the existing NAS benchmark datasets (e.g., NAS-Bench-101), but can be used to find high-performance architectures in large-scale search spaces. Particularly, we propose a progressive subspace adaptation strategy to address the domain discrepancy between the source architecture space and the target space. Considering the large difference between two architecture spaces, an assistant space is developed to smooth the transfer process. Compared with existing NAS methods, the proposed CDP is much more efficient. For example, CDP only requires the search cost of 0.1 GPU Days to find architectures with 76.9% top-1 accuracy on ImageNet and 97.51% on CIFAR-10.

## [UDC: Unified DNAS for Compressible TinyML Models for Neural Processing Units](#)

- Igor Fedorov · Ramon Matas · Hokchhay Tann · Chuteng Zhou · Matthew Mattina · Paul Whatmough
- abstract@[open-review](#): Deploying TinyML models on low-cost IoT hardware is very challenging, due to limited device memory capacity. Neural processing unit (NPU) hardware address the memory challenge by using model compression to exploit weight quantization and sparsity to fit more parameters in the same footprint. However, designing compressible neural networks (NNs) is challenging, as it expands the design space across which we must make balanced trade-offs. This paper demonstrates Unified DNAS for Compressible (UDC) NNs, which explores a large search space to generate state-of-the-art compressible NNs for NPU. ImageNet results show UDC networks are up to 3.35x smaller (iso-accuracy) or 6.25% more accurate (iso-model size) than previous work.

## [Learning Invariant Graph Representations Under Distribution Shifts](#)

- Haoyang Li · Ziwei Zhang · Xin Wang · Wenwu Zhu
- abstract@[open-review](#): Graph representation learning has shown effectiveness when testing and training graph data come from the same distribution, but most existing approaches fail to generalize under distribution shifts. Invariant learning, backed by the invariance principle from causality, can achieve guaranteed generalization under distribution shifts in theory and has shown great successes in practice. However, invariant learning for graphs under distribution shifts remains unexplored and challenging. To solve this problem, we propose Graph Invariant Learning (GIL) model capable of learning generalized graph representations under distribution shifts. Our proposed method can capture the invariant relationships between predictive graph structural information and labels in a mixture of latent environments through jointly optimizing three tailored modules. Specifically, we first design a GNN-based subgraph generator to identify invariant subgraphs. Then we use the variant subgraphs, i.e., complements of invariant subgraphs, to infer the latent environment labels. We further propose an invariant learning module to learn graph representations that can generalize to unknown test graphs. Theoretical justifications for our proposed method are also provided. Extensive experiments on both synthetic and real-world datasets demonstrate the superiority of our method against state-of-the-art baselines under distribution shifts for the graph classification task.

## [Social-Inverse: Inverse Decision-making of Social Contagion Management with Task Migrations](#)

- Guangmo Tong
- abstract@[open-review](#): Considering two decision-making tasks \$A\$ and \$B\$, each of which wishes to compute an effective decision \$Y\$ for a given query \$X\$, can we solve task \$B\$ by using query-decision pairs \$(X, Y)\$ of \$A\$ without knowing the latent decision-making model? Such problems, called inverse decision-making with task migrations, are of interest in that the complex and stochastic nature of real-world applications often prevents the agent from completely knowing the underlying system. In this paper, we introduce such a new problem with formal formulations and present a generic framework for addressing decision-making tasks in social contagion management. On the theory side, we present a generalization analysis for justifying the learning performance of our framework. In empirical studies, we perform a sanity check and compare the presented method with other possible learning-based and graph-based methods. We have acquired promising experimental results, confirming for the first time that it is possible to solve one decision-making task by using the solutions associated with another one.

## [AZ-whiteness test: a test for signal uncorrelation on spatio-temporal graphs](#)

- Daniele Zambon · Cesare Alippi
- abstract@[open-review](#): We present the first whiteness hypothesis test for graphs, i.e., a whiteness test for multivariate time series associated with the nodes of a dynamic graph; as such, the test represents an important model assessment tool for graph deep learning, e.g., in forecasting setups. The statistical test aims at detecting existing serial dependencies among close-in-time observations, as well as spatial dependencies among neighboring observations given the underlying graph. The proposed AZ-test can be intended as a spatio-temporal extension of traditional tests designed for system identification to graph signals. The AZ-test is versatile, allowing the underlying graph to be dynamic, changing in topology and set of nodes over time, and weighted, thus accounting for connections of different strength, as it is the case in many application scenarios like sensor and transportation networks. The asymptotic distribution of the designed test can be derived under the null hypothesis without assuming identically distributed data. We show the effectiveness of the test on both synthetic and real-world problems, and illustrate how it can be employed to assess the quality of spatio-temporal forecasting models by analyzing the prediction residuals appended to the graph stream.

## [Matryoshka Representation Learning](#)

- Aditya Kusupati · Gantavya Bhatt · Aniket Rege · Matthew Wallingford · Aditya Sinha · Vivek Ramanujan · William Howard-Snyder · Kaifeng Chen · Sham Kakade · Prateek Jain · Ali Farhadi
- abstract@[open-review](#): Learned representations are a central component in modern ML systems, serving a multitude of downstream tasks. When training such representations, it is often the case that computational and statistical constraints for each downstream task are unknown. In this context rigid, fixed capacity representations can be either over or under-accommodating to the task at hand. This leads us to ask: can we design a flexible representation that can adapt to multiple downstream tasks with varying computational resources? Our main contribution is Matryoshka Representation Learning (MRL) which encodes information at different granularities and allows a single embedding to adapt to the computational constraints of downstream tasks. MRL minimally modifies existing representation learning pipelines and imposes no additional cost during inference and deployment. MRL learns coarse-to-fine representations that are at least as accurate and rich as independently trained low-dimensional representations. The flexibility within the learned Matryoshka Representations offer: (a) up to \$\mathbf{14}\times\$ smaller embedding size for ImageNet-1K classification at the same level of accuracy; (b) up to \$\mathbf{14}\times\$ real-world speed-ups for large-scale retrieval on ImageNet-1K and 4K; and (c) up to \$\mathbf{2\%}\$ accuracy improvements for long-tail few-shot classification, all while being as robust as the original representations. Finally, we show that MRL extends seamlessly to web-scale datasets (ImageNet, JFT) across various modalities -- vision (ViT, ResNet), vision + language (ALIGN) and language (BERT). MRL code and pretrained models are open-sourced at <https://github.com/RAIVNLab/MRL>.

## [Amplifying Membership Exposure via Data Poisoning](#)

- Yufei Chen · Chao Shen · Yun Shen · Cong Wang · Yang Zhang
- abstract@[open-review](#): As in-the-wild data are increasingly involved in the training stage, machine learning applications become more susceptible to data poisoning attacks. Such attacks typically lead to test-time accuracy degradation or controlled misprediction. In this paper, we investigate the third type of exploitation of data poisoning - increasing the risks of privacy leakage of benign training samples. To this end, we demonstrate a set of data poisoning attacks to amplify the membership exposure of the targeted class. We first propose a generic dirty-label attack for supervised classification algorithms. We then propose an optimization-based clean-label attack in the transfer learning scenario, whereby the poisoning samples are correctly labeled and look

"natural" to evade human moderation. We extensively evaluate our attacks on computer vision benchmarks. Our results show that the proposed attacks can substantially increase the membership inference precision with minimum overall test-time model performance degradation.

## [A Spectral Approach to Item Response Theory](#)

- Duc Nguyen · Anderson Ye Zhang
- abstract@[open-review](#): The Rasch model is one of the most fundamental models in item response theory and has wide-ranging applications from education testing to recommendation systems. In a universe with  $n$  users and  $m$  items, the Rasch model assumes that the binary response  $X_{li} \in \{0,1\}$  of a user  $i$  with parameter  $\theta_l$  to an item  $i$  with parameter  $\beta_i$  (e.g., a user likes a movie, a student correctly solves a problem) is distributed as  $P(X_{li}=1) = 1/(1 + \exp(-(\theta_l - \beta_i)))$ . In this paper, we propose a new item estimation algorithm for this celebrated model (i.e., to estimate  $\beta^*$ ). The core of our algorithm is the computation of the stationary distribution of a Markov chain defined on an item-item graph. We complement our algorithmic contributions with finite-sample error guarantees, the first of their kind in the literature, showing that our algorithm is consistent and enjoys favorable optimality properties. We discuss practical modifications to accelerate and robustify the algorithm that practitioners can adopt. Experiments on synthetic and real-life datasets, ranging from small education testing datasets to large recommendation systems datasets show that our algorithm is scalable, accurate, and competitive with the most commonly used methods in the literature.

## [Composition Theorems for Interactive Differential Privacy](#)

- Xin Lyu
- abstract@[open-review](#): An interactive mechanism is an algorithm that stores a data set and answers adaptively chosen queries to it. The mechanism is called differentially private, if any adversary cannot distinguish whether a specific individual is in the data set by interacting with the mechanism. We study composition properties of differential privacy in concurrent compositions. In this setting, an adversary interacts with  $k$  interactive mechanisms in parallel and can interleave its queries to the mechanisms arbitrarily. Previously, Vadhan and Wang [TCC 2021] proved an optimal concurrent composition theorem for pure-differential privacy. We significantly generalize and extend their results. Namely, we prove optimal parallel composition properties for several major notions of differential privacy in the literature, including approximate DP, Renyi DP, and zero-concentrated DP. Our results demonstrate that the adversary gains no advantage by interleaving its queries to independently running mechanisms. Hence, interactivity is a feature that differential privacy grants us for free.

## [Quantized Training of Gradient Boosted Decision Trees](#)

- Yu Shi · Guolin Ke · Zhuoming Chen · Shuxin Zheng · Tie-Yan Liu
- abstract@[open-review](#): Recent years have witnessed significant success in Gradient Boosted Decision Trees (GBDT) for a wide range of machine learning applications. Generally, a consensus is agreement among GBDT's training algorithms that gradients and statistics are computed based on high-precision floating point. In this paper, we investigate an essentially important question but has been largely ignored by the previous literature - how many bits are in need for representing gradients in training GBDT? To solve this mystery, we propose to quantize all the high-precision gradients in a very simple yet effective way in the GBDT's training algorithm. Surprisingly, both our theoretical analysis and empirical studies show that the necessary precisions of gradients without hurting any performance can be quite low, e.g., 2 or 3 bits. With low-precision gradients, most arithmetic operations in GBDT training can be replaced by integer operations of 8, 16, or 32 bits. Promisingly, these findings may pave the way for much more efficient training of GBDT from several aspects: (1) speeding up the computation of gradients and histograms; (2) compressing the communication cost of high-precision statistical information during distributed training; (3) the inspiration of utilization and development of hardware architectures which well support low-precision computation. Benchmarked on CPU, GPU, and distributed clusters, we observe up to  $2\times$  speedup of our simple quantization strategy comparing with SOTA GBDT systems on extensive datasets, demonstrating the effectiveness and potential of the low-precision training of GBDT.

## [Are Two Heads the Same as One? Identifying Disparate Treatment in Fair Neural Networks](#)

- Michael Lohaus · Matthias Kleindessner · Krishnaram Kenthapadi · Francesco Locatello · Chris Russell
- abstract@[open-review](#): We show that deep networks that are trained to satisfy demographic parity fairness do so through a form of race or gender awareness, and that the more we force a network to be fair, the more accurately we can recover race or gender from the internal state of the network. Based on this observation, we investigate an alternative fairness approach: we add a second classification head to the network to explicitly predict the protected attribute (such as race or gender) alongside the original task. After training the two-headed network, we enforce demographic parity by merging the two heads, creating a network with the same architecture as the original network. We establish a close relationship between existing approaches and our approach by showing (1) that the decisions of a fair classifier are well approximated by our approach, and (2) that an unfair and optimally accurate classifier can be recovered from a fair classifier and our second head predicting the protected attribute. We use our explicit formulation to argue that the existing fairness approaches, just as ours, demonstrate disparate treatment and that they are likely to be unlawful in a wide range of scenarios under the US law.

## [MultiGuard: Provably Robust Multi-label Classification against Adversarial Examples](#)

- Jinyuan Jia · Wenjie Qu · Neil Gong
- abstract@[open-review](#): Multi-label classification, which predicts a set of labels for an input, has many applications. However, multiple recent studies showed that multi-label classification is vulnerable to adversarial examples. In particular, an attacker can manipulate the labels predicted by a multi-label classifier for an input via adding carefully crafted, human-imperceptible perturbation to it. Existing provable defenses for multi-class classification achieve sub-optimal provable robustness guarantees when generalized to multi-label classification. In this work, we propose MultiGuard, the first provably robust defense against adversarial examples to multi-label classification. Our MultiGuard leverages randomized smoothing, which is the state-of-the-art technique to build provably robust classifiers. Specifically, given an arbitrary multi-label classifier, our MultiGuard builds a smoothed multi-label classifier via adding random noise to the input. We consider isotropic Gaussian noise in this work. Our major theoretical contribution is that we show a certain number of ground truth labels of an input are provably in the set of labels predicted by our MultiGuard when the  $\ell_2$ -norm of the adversarial perturbation added to the input is bounded. Moreover, we design an algorithm to compute our provable robustness guarantees. Empirically, we evaluate our MultiGuard on VOC 2007, MS-COCO, and NUS-WIDE benchmark datasets. Our code is available at: <https://github.com/quwenjie/MultiGuard>

## [Graph Scattering beyond Wavelet Shackles](#)

- Christian Koke · Gitta Kutyniok
- abstract@[open-review](#): This work develops a flexible and mathematically sound framework for the design and analysis of graph scattering networks with variable branching ratios and generic functional calculus filters. Spectrally-agnostic stability guarantees for node- and graph-level perturbations are derived; the vertex-set non-preserving case is treated by utilizing recently developed mathematical-physics based tools. Energy propagation through the network layers is investigated and related to truncation stability. New methods of graph-level feature aggregation are introduced and stability of the resulting composite scattering architectures is established. Finally, scattering transforms are extended to edge- and higher order tensorial input. Theoretical results are complemented by numerical investigations: Suitably chosen scattering networks conforming to the developed theory perform better than traditional graph-wavelet based scattering approaches in social network graph classification tasks and significantly outperform other graph-based learning approaches to regression of quantum-chemical energies on QM7\$.

## Relational Reasoning via Set Transformers: Provable Efficiency and Applications to MARL

- Fengzhuo Zhang · Boyi Liu · KAXIN WANG · Vincent Tan · Zhuoran Yang · Zhaoran Wang
- abstract@[open-review](#): The cooperative Multi-Agent Reinforcement Learning (MARL) with permutation invariant agents framework has achieved tremendous empirical successes in real-world applications. Unfortunately, the theoretical understanding of this MARL problem is lacking due to the curse of many agents and the limited exploration of the relational reasoning in existing works. In this paper, we verify that the transformer implements complex relational reasoning, and we propose and analyze model-free and model-based offline MARL algorithms with the transformer approximators. We prove that the suboptimality gaps of the model-free and model-based algorithms are independent of and logarithmic in the number of agents respectively, which mitigates the curse of many agents. These results are consequences of a novel generalization error bound of the transformer and a novel analysis of the Maximum Likelihood Estimate (MLE) of the system dynamics with the transformer. Our model-based algorithm is the first provably efficient MARL algorithm that explicitly exploits the permutation invariance of the agents.

## Causal Inference with Non-IID Data using Linear Graphical Models

- Chi Zhang · Karthika Mohan · Judea Pearl
- abstract@[open-review](#): Traditional causal inference techniques assume data are independent and identically distributed (IID) and thus ignores interactions among units. However, a unit's treatment may affect another unit's outcome (interference), a unit's treatment may be correlated with another unit's outcome or a unit's treatment and outcome may be spuriously correlated through another unit. To capture such nuances, we model the data generating process using causal graphs and conduct a systematic analysis of the bias caused by different types of interactions when computing causal effects. We derive theorems to detect and quantify the interaction bias, and derive conditions under which it is safe to ignore interactions. Put differently, we present conditions under which causal effects can be computed with negligible bias by assuming that samples are IID. Furthermore, we develop a method to eliminate bias in cases where blindly assuming IID is expected to yield a significantly biased estimate. Finally, we test the coverage and performance of our methods through simulations.

## A Unifying Framework of Off-Policy General Value Function Evaluation

- Tengyu Xu · Zhuoran Yang · Zhaoran Wang · Yingbin Liang
- abstract@[open-review](#): General Value Function (GVF) is a powerful tool to represent both the {\em predictive} and {\em retrospective} knowledge in reinforcement learning (RL). In practice, often multiple interrelated GVF need to be evaluated jointly with pre-collected off-policy samples. In the literature, the gradient temporal difference (GTD) learning method has been adopted to evaluate GVF in the off-policy setting, but such an approach may suffer from a large estimation error even if the function approximation class is sufficiently expressive. Moreover, none of the previous work have formally established the convergence guarantee to the ground truth GVF under the function approximation settings. In this paper, we address both issues through the lens of a class of GVF with causal filtering, which cover a wide range of RL applications such as reward variance, value gradient, cost in anomaly detection, stationary distribution gradient, etc. We propose a new algorithm called GenTD for off-policy GVF evaluation and show that GenTD learns multiple interrelated multi-dimensional GVF as efficiently as a single canonical scalar value function. We further show that unlike GTD, the learned GVF by GenTD are guaranteed to converge to the ground truth GVF as long as the function approximation power is sufficiently large. To our best knowledge, GenTD is the first off-policy GVF evaluation algorithm that has global optimality guarantee.

## Make an Omelette with Breaking Eggs: Zero-Shot Learning for Novel Attribute Synthesis

- Yu-Hsuan Li · Tzu-Yin Chao · Ching-Chun Huang · Pin-Yu Chen · Wei-Chen Chiu
- abstract@[open-review](#): Most of the existing algorithms for zero-shot classification problems typically rely on the attribute-based semantic relations among categories to realize the classification of novel categories without observing any of their instances. However, training the zero-shot classification models still requires attribute labeling for each class (or even instance) in the training dataset, which is also expensive. To this end, in this paper, we bring up a new problem scenario: "Can we derive zero-shot learning for novel attribute detectors/classifiers and use them to automatically annotate the dataset for labeling efficiency?" Basically, given only a small set of detectors that are learned to recognize some manually annotated attributes (i.e., the seen attributes), we aim to synthesize the detectors of novel attributes in a zero-shot learning manner. Our proposed method, Zero-Shot Learning for Attributes (ZSLA), which is the first of its kind to the best of our knowledge, tackles this new research problem by applying the set operations to first decompose the seen attributes into their basic attributes and then recombine these basic attributes into the novel ones. Extensive experiments are conducted to verify the capacity of our synthesized detectors for accurately capturing the semantics of the novel attributes and show their superior performance in terms of detection and localization compared to other baseline approaches. Moreover, we demonstrate the application of automatic annotation using our synthesized detectors on Caltech-UCSD Birds-200-2011 dataset. Various generalized zero-shot classification algorithms trained upon the dataset re-annotated by ZSLA shows comparable performance with those trained with the manual ground-truth annotations.

## Exact learning dynamics of deep linear networks with prior knowledge

- Lukas Braun · Clémentine Domini · James Fitzgerald · Andrew Saxe
- abstract@[open-review](#): Learning in deep neural networks is known to depend critically on the knowledge embedded in the initial network weights. However, few theoretical results have precisely linked prior knowledge to learning dynamics. Here we derive exact solutions to the dynamics of learning with rich prior knowledge in deep linear networks by generalising Fukumizu's matrix Riccati solution \citet{Fukumizu1998}. While simple, deep linear networks retain a non-convex loss landscape and nonlinear learning dynamics that depend in detail on the initial weights of the network. We obtain explicit expressions for the evolving network function, hidden representational similarity, and neural tangent kernel over training for a broad class of initialisations and tasks. We characterise a class of task-independent initialisations that radically alters learning dynamics from slow step-like to fast exponential trajectories while converging to identical representational similarity, dissociating learning trajectories from the structure of internal representations. We discuss the implications of this finding for neural network weight initialisation schemes, continual learning and learning of structured knowledge. Finally, we characterise how network weights dynamically align with task structure, rigorously justifying why previous solutions successfully described learning from small weights without incorporating their fine-scale structure. Taken together, our results provide a mathematical toolkit for understanding the impact of prior knowledge on deep learning.

## Top Two Algorithms Revisited

- Marc Jourdan · Remy Degenne · Dorian Baudry · Rianne de Heide · Emilie Kaufmann
- abstract@[open-review](#): Top two algorithms arose as an adaptation of Thompson sampling to best arm identification in multi-armed bandit models for parametric families of arms. They select the next arm to sample from by randomizing among two candidate arms, a leader and a challenger. Despite their good empirical performance, theoretical guarantees for fixed-confidence best arm identification have only been obtained when the arms are Gaussian with known variances. In this paper, we provide a general analysis of top-two methods, which identifies desirable properties of the leader, the challenger, and the (possibly non-parametric) distributions of the arms. As a result, we obtain theoretically supported top-two algorithms for best arm identification with bounded distributions. Our proof method demonstrates in particular that the sampling step used to select the leader inherited from Thompson sampling can be replaced by other choices, like selecting the empirical best arm.

## Robust Reinforcement Learning using Offline Data

- Kishan Panaganti · Zaiyan Xu · Dileep Kalathil · Mohammad Ghavamzadeh
- abstract@[open-review](#): The goal of robust reinforcement learning (RL) is to learn a policy that is robust against the uncertainty in model parameters. Parameter uncertainty commonly occurs in many real-world RL applications due to simulator modeling errors, changes in the real-world system dynamics over time, and adversarial disturbances. Robust RL is typically formulated as a max-min problem, where the objective is to learn the policy that maximizes the value against the worst possible models that lie in an uncertainty set. In this work, we propose a robust RL algorithm called Robust Fitted Q-Iteration (RFQI), which uses only an offline dataset to learn the optimal robust policy. Robust RL with offline data is significantly more challenging than its non-robust counterpart because of the minimization over all models present in the robust Bellman operator. This poses challenges in offline data collection, optimization over the models, and unbiased estimation. In this work, we propose a systematic approach to overcome these challenges, resulting in our RFQI algorithm. We prove that RFQI learns a near-optimal robust policy under standard assumptions and demonstrate its superior performance on standard benchmark problems.

## [SnAKE: Bayesian Optimization with Pathwise Exploration](#)

- Jose Pablo Folch · Shiqiang Zhang · Robert Lee · Behrang Shafei · David Walz · Calvin Tsay · Mark van der Wilk · Ruth Misener
- abstract@[open-review](#): "Bayesian Optimization is a very effective tool for optimizing expensive black-box functions. Inspired by applications developing and characterizing reaction chemistry using droplet microfluidic reactors, we consider a novel setting where the expense of evaluating the function can increase significantly when making large input changes between iterations. We further assume we are working asynchronously, meaning we have to decide on new queries before we finish evaluating previous experiments. This paper investigates the problem and introduces 'Sequential Bayesian Optimization via Adaptive Connecting Samples' (SnAKE), which provides a solution by considering large batches of queries and preemptively building optimization paths that minimize input costs. We investigate some convergence properties and empirically show that the algorithm is able to achieve regret similar to classical Bayesian Optimization algorithms in both the synchronous and asynchronous settings, while reducing the input costs significantly. We show the method is robust to the choice of its single hyper-parameter and provide a parameter-free alternative."

## [Turbocharging Solution Concepts: Solving NEs, CEs and CCEs with Neural Equilibrium Solvers](#)

- Luke Marris · Ian Gemp · Thomas Anthony · Andrea Tacchetti · Siqi Liu · Karl Tuyls
- abstract@[open-review](#): Solution concepts such as Nash Equilibria, Correlated Equilibria, and Coarse Correlated Equilibria are useful components for many multiagent machine learning algorithms. Unfortunately, solving a normal-form game could take prohibitive or non-deterministic time to converge, and could fail. We introduce the Neural Equilibrium Solver which utilizes a special equivariant neural network architecture to approximately solve the space of all games of fixed shape, buying speed and determinism. We define a flexible equilibrium selection framework, that is capable of uniquely selecting an equilibrium that minimizes relative entropy, or maximizes welfare. The network is trained without needing to generate any supervised training data. We show remarkable zero-shot generalization to larger games. We argue that such a network is a powerful module for many possible multiagent algorithms.

## [Fast Stochastic Composite Minimization and an Accelerated Frank-Wolfe Algorithm under Parallelization](#)

- Benjamin Dubois-Taine · Francis Bach · Quentin Berthet · Adrien Taylor
- abstract@[open-review](#): We consider the problem of minimizing the sum of two convex functions. One of those functions has Lipschitz-continuous gradients, and can be accessed via stochastic oracles, whereas the other is ``simple''. We provide a Bregman-type algorithm with accelerated convergence in function values to a ball containing the minimum. The radius of this ball depends on problem-dependent constants, including the variance of the stochastic oracle. We further show that this algorithmic setup naturally leads to a variant of Frank-Wolfe achieving acceleration under parallelization. More precisely, when minimizing a smooth convex function on a bounded domain, we show that one can achieve an  $\tilde{O}(1/\sqrt{\epsilon})$  primal-dual gap (in expectation) in  $\tilde{O}(1/\sqrt{\epsilon})$  iterations, by only accessing gradients of the original function and a linear maximization oracle with  $O(1/\sqrt{\epsilon})$  computing units in parallel. We illustrate this fast convergence on synthetic numerical experiments.

## [On Viewpoint Robustness of Visual Recognition in the Wild](#)

- Yinpeng Dong · Shouwei Ruan · Hang Su · Caixin Kang · Xingxing Wei · Jun Zhu
- abstract@[open-review](#): Recent studies have demonstrated that visual recognition models lack robustness to distribution shift. However, current work mainly considers model robustness to 2D image transformations, leaving viewpoint changes in the 3D world less explored. In general, viewpoint changes are prevalent in various real-world applications (e.g., autonomous driving), making it imperative to evaluate viewpoint robustness. In this paper, we propose a novel method called ViewFool to find adversarial viewpoints that mislead visual recognition models. By encoding real-world objects as neural radiance fields (NeRF), ViewFool characterizes a distribution of diverse adversarial viewpoints under an entropic regularizer, which helps to handle the fluctuations of the real camera pose and mitigate the reality gap between the real objects and their neural representations. Experiments validate that the common image classifiers are extremely vulnerable to the generated adversarial viewpoints, which also exhibit high cross-model transferability. Based on ViewFool, we introduce ImageNet-V, a new out-of-distribution dataset for benchmarking viewpoint robustness of image classifiers. Evaluation results on 40 classifiers with diverse architectures, objective functions, and data augmentations reveal a significant drop in model performance when tested on ImageNet-V, which provides a possibility to leverage ViewFool as an effective data augmentation strategy to improve viewpoint robustness.

## [Fairness Transferability Subject to Bounded Distribution Shift](#)

- Yatong Chen · Reilly Raab · Jialu Wang · Yang Liu
- abstract@[open-review](#): Given an algorithmic predictor that is "fair" on some source distribution, will it still be fair on an unknown target distribution that differs from the source within some bound? In this paper, we study the transferability of statistical group fairness for machine learning predictors (i.e., classifiers or regressors) subject to bounded distribution shift, a phenomenon frequently caused by user adaptation to a deployed model or a dynamic environment. Herein, we develop a bound characterizing such transferability, flagging potentially inappropriate deployments of machine learning for socially consequential tasks. We first develop a framework for bounding violations of statistical fairness subject to distribution shift, formulating a generic upper bound for transferred fairness violation as our primary result. We then develop bounds for specific worked examples, adopting two commonly used fairness definitions (i.e., demographic parity and equalized odds) for two classes of distribution shift (i.e., covariate shift and label shift). Finally, we compare our theoretical bounds to deterministic models of distribution shift and against real-world data, finding that we are able to estimate fairness violation bounds in practice, even when simplifying assumptions are only approximately satisfied.

## [SelecMix: Debiased Learning by Contradicting-pair Sampling](#)

- Inwoo Hwang · Sangjun Lee · Yunhyeok Kwak · Seong Joon Oh · Damien Teney · Jin-Hwa Kim · Byoung-Tak Zhang
- abstract@[open-review](#): Neural networks trained with ERM (empirical risk minimization) sometimes learn unintended decision rules, in particular when their training data is biased, i.e., when training labels are correlated with undesirable features. Techniques have been proposed to prevent a network from learning such features, using the heuristic that spurious correlations are ``simple'' and learned preferentially during training. Recent methods augment training data such that samples displaying spurious correlations (i.e., bias-aligned samples) become a minority, whereas the other, bias-conflicting samples become prevalent. However, they require sophisticated techniques such as disentanglement with careful tuning, making them challenging to train the model and scale to complex real-world datasets. In this work, we use the mixup, which builds convex combinations of input images and their labels, to augment the bias-conflicting samples. Mainly, we propose a selective mixup scheme, SelecMix, where the mixup is applied to the pairs having (i) the

same label but dissimilar biased features, and (ii) different labels but similar biased features. To compare samples with respect to the biased features, we propose the bias-amplified contrastive model relying on the heuristic that biased features are learned preferentially during training. Experimental results demonstrate the effectiveness of the proposed method on both the synthetic and real-world datasets. Furthermore, we validate the robustness of our method under noisy labels, which is more realistic, and challenging to identify bias-conflicting samples.

## [Inducing Equilibria via Incentives: Simultaneous Design-and-Play Ensures Global Convergence](#)

- Boyi Liu · Jiayang Li · Zhuoran Yang · Hoi-To Wai · Mingyi Hong · Yu Nie · Zhaoran Wang
- abstract@[open-review](#): To regulate a social system comprised of self-interested agents, economic incentives are often required to induce a desirable outcome. This incentive design problem naturally possesses a bilevel structure, in which a designer modifies the payoffs of the agents with incentives while anticipating the response of the agents, who play a non-cooperative game that converges to an equilibrium. The existing bilevel optimization algorithms raise a dilemma when applied to this problem: anticipating how incentives affect the agents at equilibrium requires solving the equilibrium problem repeatedly, which is computationally inefficient; bypassing the time-consuming step of equilibrium-finding can reduce the computational cost, but may lead the designer to a sub-optimal solution. To address such a dilemma, we propose a method that tackles the designer's and agents' problems simultaneously in a single loop. Specifically, at each iteration, both the designer and the agents only move one step. Nevertheless, we allow the designer to gradually learn the overall influence of the incentives on the agents, which guarantees optimality after convergence. The convergence rate of the proposed scheme is also established for a broad class of games.

## [A composable machine-learning approach for steady-state simulations on high-resolution grids](#)

- Rishikesh Ranade · Chris Hill · Lalit Ghule · Jay Pathak
- abstract@[open-review](#): In this paper we show that our Machine Learning (ML) approach, CoMLSim (Composable Machine Learning Simulator), can simulate PDEs on highly-resolved grids with higher accuracy and generalization to out-of-distribution source terms and geometries than traditional ML baselines. Our unique approach combines key principles of traditional PDE solvers with local-learning and low-dimensional manifold techniques to iteratively simulate PDEs on large computational domains. The proposed approach is validated on more than 5 steady-state PDEs across different PDE conditions on highly-resolved grids and comparisons are made with the commercial solver, Ansys Fluent as well 4 other state-of-the-art ML methods. The numerical experiments show that our approach outperforms ML baselines in terms of 1) accuracy across quantitative metrics and 2) generalization to out-of-distribution conditions as well as mesh resolutions. Additionally, we provide results of conducting a large number of ablations experiments to highlight components of our approach that strongly influence the results. We conclude that our local-learning and iterative-inferencing approach reduces the challenge of generalization that most ML models face.

## [Multi-agent Dynamic Algorithm Configuration](#)

- Ke Xue · Jiacheng Xu · Lei Yuan · Miqing Li · Chao Qian · Zongzhang Zhang · Yang Yu
- abstract@[open-review](#): Automated algorithm configuration relieves users from tedious, trial-and-error tuning tasks. A popular algorithm configuration tuning paradigm is dynamic algorithm configuration (DAC), in which an agent learns dynamic configuration policies across instances by reinforcement learning (RL). However, in many complex algorithms, there may exist different types of configuration hyperparameters, and such heterogeneity may bring difficulties for classic DAC which uses a single-agent RL policy. In this paper, we aim to address this issue and propose a multi-agent DAC (MA-DAC), with one agent working for one type of configuration hyperparameters. MA-DAC formulates the dynamic configuration of a complex algorithm with heterogeneous types of hyperparameters as a contextual multi-agent Markov decision process and solves it by a cooperative multi-agent RL (MARL) algorithm. To instantiate, we apply MA-DAC to a well-known optimization algorithm for multi-objective optimization problems. Experimental results show the effectiveness of MA-DAC in not only achieving superior performance compared with other configuration tuning approaches based on heuristic rules, multi-armed bandits, and single-agent RL, but also being capable of generalizing to different problem classes. Furthermore, we release the environments in this paper as a benchmark for testing MARL algorithms, with the hope of facilitating the application of MARL.

## [Contrastive and Non-Contrastive Self-Supervised Learning Recover Global and Local Spectral Embedding Methods](#)

- Randall Balestrieri · Yann LeCun
- abstract@[open-review](#): Self-Supervised Learning (SSL) surmises that inputs and pairwise positive relationships are enough to learn meaningful representations. Although SSL has recently reached a milestone: outperforming supervised methods in many modalities\theoretical foundations are limited, method-specific, and fail to provide principled design guidelines to practitioners. In this paper, we propose a unifying framework under the helm of spectral manifold learning. Through the course of this study, we will demonstrate that VICReg, SimCLR, BarlowTwins et al. correspond to eponymous spectral methods such as Laplacian Eigenmaps, ISOMAP et al. From this unified viewpoint, we obtain (i) the close-form optimal representation, (ii) the close-form optimal network parameters in the linear regime, (iii) the impact of the pairwise relations used during training on each of those quantities and on downstream task performances, and most importantly, (iv) the first theoretical bridge between contrastive and non-contrastive methods to global and local spectral methods respectively hinting at the benefits and limitations of each. For example, if the pairwise relation is aligned with the downstream task, all SSL methods produce optimal representations for that downstream task.

## [On the Double Descent of Random Features Models Trained with SGD](#)

- Fanghui Liu · Johan Suykens · Volkan Cevher
- abstract@[open-review](#): We study generalization properties of random features (RF) regression in high dimensions optimized by stochastic gradient descent (SGD) in under-/over-parameterized regime. In this work, we derive precise non-asymptotic error bounds of RF regression under both constant and polynomial-decay step-size SGD setting, and observe the double descent phenomenon both theoretically and empirically. Our analysis shows how to cope with multiple randomness sources of initialization, label noise, and data sampling (as well as stochastic gradients) with no closed-form solution, and also goes beyond the commonly-used Gaussian/spherical data assumption. Our theoretical results demonstrate that, with SGD training, RF regression still generalizes well for interpolation learning, and is able to characterize the double descent behavior by the unimodality of variance and monotonic decrease of bias. Besides, we also prove that the constant step-size SGD setting incurs no loss in convergence rate when compared to the exact minimum-norm interpolator, as a theoretical justification of using SGD in practice.

## [Symplectic Spectrum Gaussian Processes: Learning Hamiltonians from Noisy and Sparse Data](#)

- Yusuke Tanaka · Tomoharu Iwata · naonori ueda
- abstract@[open-review](#): Hamiltonian mechanics is a well-established theory for modeling the time evolution of systems with conserved quantities (called Hamiltonian), such as the total energy of the system. Recent works have parameterized the Hamiltonian by machine learning models (e.g., neural networks), allowing Hamiltonian dynamics to be obtained from state trajectories without explicit mathematical modeling. However, the performance of existing models is limited as we can observe only noisy and sparse trajectories in practice. This paper proposes a probabilistic model that can learn the dynamics of conservative or dissipative systems from noisy and sparse data. We introduce a Gaussian process that incorporates the geometric structure (called symplectic structure) of Hamiltonian mechanics, which is used as a prior distribution for estimating Hamiltonian systems with additive dissipation. We then introduce its spectral representation, Symplectic Spectrum Gaussian Processes (SSGPs), for which we newly derive random Fourier features with symplectic structures. This allows us to construct an efficient variational inference algorithm for training the models while simulating the dynamics via

ordinary differential equation solvers. Experiments on several physical systems show that SSGP offers excellent performance in predicting dynamics that follow the energy conservation or dissipation law from noisy and sparse data.

## [Learning dynamics of deep linear networks with multiple pathways](#)

- Jianghong Shi · Eric Shea-Brown · Michael Buice
- abstract@[open-review](#): Not only have deep networks become standard in machine learning, they are increasingly of interest in neuroscience as models of cortical computation that capture relationships between structural and functional properties. In addition they are a useful target of theoretical research into the properties of network computation. Deep networks typically have a serial or approximately serial organization across layers, and this is often mirrored in models that purport to represent computation in mammalian brains. There are, however, multiple examples of parallel pathways in mammalian brains. In some cases, such as the mouse, the entire visual system appears arranged in a largely parallel, rather than serial fashion. While these pathways may be formed by differing cost functions that drive different computations, here we present a new mathematical analysis of learning dynamics in networks that have parallel computational pathways driven by the same cost function. We use the approximation of deep linear networks with large hidden layer sizes to show that, as the depth of the parallel pathways increases, different features of the training set (defined by the singular values of the input-output correlation) will typically concentrate in one of the pathways. This result is derived analytically and demonstrated with numerical simulation. Thus, rather than sharing stimulus and task features across multiple pathways, parallel network architectures learn to produce sharply diversified representations with specialized and specific pathways, a mechanism which may hold important consequences for codes in both biological and artificial systems.

## [VoxGRAF: Fast 3D-Aware Image Synthesis with Sparse Voxel Grids](#)

- Katja Schwarz · Axel Sauer · Michael Niemeyer · Yiyi Liao · Andreas Geiger
- abstract@[open-review](#): State-of-the-art 3D-aware generative models rely on coordinate-based MLPs to parameterize 3D radiance fields. While demonstrating impressive results, querying an MLP for every sample along each ray leads to slow rendering. Therefore, existing approaches often render low-resolution feature maps and process them with an upsampling network to obtain the final image. Albeit efficient, neural rendering often entangles viewpoint and content such that changing the camera pose results in unwanted changes of geometry or appearance. Motivated by recent results in voxel-based novel view synthesis, we investigate the utility of sparse voxel grid representations for fast and 3D-consistent generative modeling in this paper. Our results demonstrate that monolithic MLPs can indeed be replaced by 3D convolutions when combining sparse voxel grids with progressive growing, free space pruning and appropriate regularization. To obtain a compact representation of the scene and allow for scaling to higher voxel resolutions, our model disentangles the foreground object (modeled in 3D) from the background (modeled in 2D). In contrast to existing approaches, our method requires only a single forward pass to generate a full 3D scene. It hence allows for efficient rendering from arbitrary viewpoints while yielding 3D consistent results with high visual fidelity.

## [Sampling with Riemannian Hamiltonian Monte Carlo in a Constrained Space](#)

- Yunbum Kook · Yin-Tat Lee · Ruqi Shen · Santosh Vempala
- abstract@[open-review](#): We demonstrate for the first time that ill-conditioned, non-smooth, constrained distributions in very high dimension, upwards of 100,000, can be sampled efficiently \emph{in practice}. Our algorithm incorporates constraints into the Riemannian version of Hamiltonian Monte Carlo and maintains sparsity. This allows us to achieve a mixing rate independent of smoothness and condition numbers. On benchmark data sets in systems biology and linear programming, our algorithm outperforms existing packages by orders of magnitude. In particular, we achieve a 1,000-fold speed-up for sampling from the largest published human metabolic network (RECON3D). Our package has been incorporated into a popular Bioinformatics library.

## [Continual learning: a feature extraction formalization, an efficient algorithm, and fundamental obstructions](#)

- Binghui Peng · Andrej Risteski
- abstract@[open-review](#): Continual learning is an emerging paradigm in machine learning, wherein a model is exposed in an online fashion to data from multiple different distributions (i.e. environments), and is expected to adapt to the distribution change. Precisely, the goal is to perform well in the new environment, while simultaneously retaining the performance on the previous environments (i.e. avoid ``catastrophic forgetting''). While this setup has enjoyed a lot of attention in the applied community, there hasn't been theoretical work that even formalizes the desired guarantees. In this paper, we propose a framework for continual learning through the framework of feature extraction---namely, one in which features, as well as a classifier, are being trained with each environment. When the features are linear, we design an efficient gradient-based algorithm  $\text{DPGrad}$ , that is guaranteed to perform well on the current environment, as well as avoid catastrophic forgetting. In the general case, when the features are non-linear, we show such an algorithm cannot exist, whether efficient or not.

## [Lottery Tickets on a Data Diet: Finding Initializations with Sparse Trainable Networks](#)

- Mansheej Paul · Brett Larsen · Surya Ganguli · Jonathan Frankle · Gintare Karolina Dziugaite
- abstract@[open-review](#): A striking observation about iterative magnitude pruning (IMP; Frankle et al. 2020) is that after just a few hundred steps of dense training—the method can find a sparse sub-network that can be trained to the same accuracy as the dense network. However, the same does not hold at step 0, i.e. random initialization. In this work, we seek to understand how this early phase of pre-training leads to a good initialization for IMP both through the lens of the data distribution and the loss landscape geometry. Empirically we observe that, holding the number of pre-training iterations constant, training on a small fraction of (randomly chosen) data suffices to obtain an equally good initialization for IMP. We additionally observe that by pre-training only on "easy" training data, we can decrease the number of steps necessary to find a good initialization for IMP compared to training on the full dataset or a randomly chosen subset. Finally, we identify novel properties of the loss landscape of dense networks that are predictive of IMP performance, showing in particular that more examples being linearly mode connected in the dense network correlates well with good initializations for IMP. Combined, these results provide new insight into the role played by the early phase training in IMP.

## [Equivariant Graph Hierarchy-based Neural Networks](#)

- Jiaqi Han · Yu Rong · Tingyang Xu · Wenbing Huang
- abstract@[open-review](#): Equivariant Graph neural Networks (EGNs) are powerful in characterizing the dynamics of multi-body physical systems. Existing EGNs conduct flat message passing, which, yet, is unable to capture the spatial/dynamical hierarchy for complex systems particularly, limiting substructure discovery and global information fusion. In this paper, we propose Equivariant Hierarchy-based Graph Networks (EGHNs) which consist of the three key components: generalized Equivariant Matrix Message Passing (EMMP) , E-Pool and E-UnPool. In particular, EMMP is able to improve the expressivity of conventional equivariant message passing, E-Pool assigns the quantities of the low-level nodes into high-level clusters, while E-UnPool leverages the high-level information to update the dynamics of the low-level nodes. As their names imply, both E-Pool and E-UnPool are guaranteed to be equivariant to meet physic symmetry. Considerable experimental evaluations verify the effectiveness of our EGHN on several applications including multi-object dynamics simulation, motion capture, and protein dynamics modeling.

## [Mirror Descent Maximizes Generalized Margin and Can Be Implemented Efficiently](#)

- Haoyuan Sun · Kwangjun Ahn · Christos Thrampoulidis · Navid Azizan

- abstract@[open-review](#): Driven by the empirical success and wide use of deep neural networks, understanding the generalization performance of overparameterized models has become an increasingly popular question. To this end, there has been substantial effort to characterize the implicit bias of the optimization algorithms used, such as gradient descent (GD), and the structural properties of their preferred solutions. This paper answers an open question in this literature: For the classification setting, what solution does mirror descent (MD) converge to? Specifically, motivated by its efficient implementation, we consider the family of mirror descent algorithms with potential function chosen as the  $\$p\$$ -th power of the  $\|\cdot\|_p$ -norm, which is an important generalization of GD. We call this algorithm  $\$p\$$ -textsf{GD}. For this family, we characterize the solutions it obtains and show that it converges in direction to a generalized maximum-margin solution with respect to the  $\|\cdot\|_p$ -norm for linearly separable classification. While the MD update rule is in general expensive to compute and not suitable for deep learning,  $\$p\$$ -textsf{GD} is fully parallelizable in the same manner as SGD and can be used to train deep neural networks with virtually no additional computational overhead. Using comprehensive experiments with both linear and deep neural network models, we demonstrate that  $\$p\$$ -textsf{GD} can noticeably affect the structure and the generalization performance of the learned models.

## [Recall Distortion in Neural Network Pruning and the Undecayed Pruning Algorithm](#)

- Aidan Good · Jiaqi Lin · Hannah Sieg · Mikey Ferguson · Xin Yu · Shandian Zhe · Jerzy Wieczorek · Thiago Serra
- abstract@[open-review](#): Pruning techniques have been successfully used in neural networks to trade accuracy for sparsity. However, the impact of network pruning is not uniform: prior work has shown that the recall for underrepresented classes in a dataset may be more negatively affected. In this work, we study such relative distortions in recall by hypothesizing an intensification effect that is inherent to the model. Namely, that pruning makes recall relatively worse for a class with recall below accuracy and, conversely, that it makes recall relatively better for a class with recall above accuracy. In addition, we propose a new pruning algorithm aimed at attenuating such effect. Through statistical analysis, we have observed that intensification is less severe with our algorithm but nevertheless more pronounced with relatively more difficult tasks, less complex models, and higher pruning ratios. More surprisingly, we conversely observe a de-intensification effect with lower pruning ratios.

## [TaSIL: Taylor Series Imitation Learning](#)

- Daniel Pfrommer · Thomas Zhang · Stephen Tu · Nikolai Matni
- abstract@[open-review](#): We propose Taylor Series Imitation Learning (TaSIL), a simple augmentation to standard behavior cloning losses in the context of continuous control. TaSIL penalizes deviations in the higher-order Taylor series terms between the learned and expert policies. We show that experts satisfying a notion of incremental input-to-state stability are easy to learn, in the sense that a small TaSIL-augmented imitation loss over expert trajectories guarantees a small imitation loss over trajectories generated by the learned policy. We provide sample-complexity bounds for TaSIL that scale as  $\tilde{\mathcal{O}}(1/n)$  in the realizable setting, for  $n$  the number of expert demonstrations. Finally, we demonstrate experimentally the relationship between the robustness of the expert policy and the order of Taylor expansion required in TaSIL, and compare standard Behavior Cloning, DART, and DAgger with TaSIL-loss-augmented variants. In all cases, we show significant improvement over baselines across a variety of MuJoCo tasks.

## [FourierFormer: Transformer Meets Generalized Fourier Integral Theorem](#)

- Tan Nguyen · Minh Pham · Tam Nguyen · Khai Nguyen · Stanley Osher · Nhat Ho
- abstract@[open-review](#): Multi-head attention empowers the recent success of transformers, the state-of-the-art models that have achieved remarkable success in sequence modeling and beyond. These attention mechanisms compute the pairwise dot products between the queries and keys, which results from the use of unnormalized Gaussian kernels with the assumption that the queries follow a mixture of Gaussian distribution. There is no guarantee that this assumption is valid in practice. In response, we first interpret attention in transformers as a nonparametric kernel regression. We then propose the FourierFormer, a new class of transformers in which the dot-product kernels are replaced by the novel generalized Fourier integral kernels. Different from the dot-product kernels, where we need to choose a good covariance matrix to capture the dependency of the features of data, the generalized Fourier integral kernels can automatically capture such dependency and remove the need to tune the covariance matrix. We theoretically prove that our proposed Fourier integral kernels can efficiently approximate any key and query distributions. Compared to the conventional transformers with dot-product attention, FourierFormers attain better accuracy and reduce the redundancy between attention heads. We empirically corroborate the advantages of FourierFormers over the baseline transformers in a variety of practical applications including language modeling and image classification.

## [Fast Bayesian Inference of Point Process Intensity as Function of Covariates](#)

- Hideaki Kim · Taichi Asami · Hiroyuki Toda
- abstract@[open-review](#): In this paper, we tackle the Bayesian inferencing of point process intensity as a function of covariates. We propose a novel augmentation of permanental process called { $\tilde{\pi}$ it augmented permanental process}, a double stochastic point process that uses a Gaussian process on covariate space to describe the Bayesian a priori uncertainty present in the square root of intensity, and derive a fast Bayesian inference algorithm that scales linearly with data size without relying on either domain discretization or Markov Chain Monte Carlo computation. The proposed algorithm is based on our finding that the representer theorem, which is related to RKHS theory, holds for augmented permanental process, which provides us with many significant computational advantages. We evaluate our algorithm on synthetic and real-world data, and show that it outperforms state-of-the-art methods in terms of predictive accuracy while being substantially faster than a conventional Bayesian method.

## [Posterior and Computational Uncertainty in Gaussian Processes](#)

- Jonathan Wenger · Geoff Pleiss · Marvin Pförtner · Philipp Hennig · John Cunningham
- abstract@[open-review](#): Gaussian processes scale prohibitively with the size of the dataset. In response, many approximation methods have been developed, which inevitably introduce approximation error. This additional source of uncertainty, due to limited computation, is entirely ignored when using the approximate posterior. Therefore in practice, GP models are often as much about the approximation method as they are about the data. Here, we develop a new class of methods that provides consistent estimation of the combined uncertainty arising from both the finite number of data observed and the finite amount of computation expended. The most common GP approximations map to an instance in this class, such as methods based on the Cholesky factorization, conjugate gradients, and inducing points. For any method in this class, we prove (i) convergence of its posterior mean in the associated RKHS, (ii) decomposability of its combined posterior covariance into mathematical and computational covariances, and (iii) that the combined variance is a tight worst-case bound for the squared error between the method's posterior mean and the latent function. Finally, we empirically demonstrate the consequences of ignoring computational uncertainty and show how it improves generalization performance on benchmark datasets.

## [Robust Streaming PCA](#)

- Daniel Bienstock · Minchan Jeong · Apurv Shukla · Se-Young Yun
- abstract@[open-review](#): We consider streaming principal component analysis (PCA) when the stochastic data-generating model is subject to perturbations. While existing models assume a fixed covariance, we adopt a robust perspective where the covariance matrix belongs to a temporal uncertainty set. Under this setting, we provide fundamental limits on any algorithm recovering principal components. We analyze the convergence of the noisy power method and Ojaâ€™s algorithm, both analyzed for the stationary data generating model, and argue that the noisy power method is rate-optimal in our setting. Finally, we demonstrate the validity of our analysis through numerical experiments.

## Efficient Graph Similarity Computation with Alignment Regularization

- Wei Zhuo · Guang Tan
- abstract@[open-review](#): We consider the graph similarity computation (GSC) task based on graph edit distance (GED) estimation. State-of-the-art methods treat GSC as a learning-based prediction task using Graph Neural Networks (GNNs). To capture fine-grained interactions between pair-wise graphs, these methods mostly contain a node-level matching module in the end-to-end learning pipeline, which causes high computational costs in both the training and inference stages. We show that the expensive node-to-node matching module is not necessary for GSC, and high-quality learning can be attained with a simple yet powerful regularization technique, which we call the Alignment Regularization (AReg). In the training stage, the AReg term imposes a node-graph correspondence constraint on the GNN encoder. In the inference stage, the graph-level representations learned by the GNN encoder are directly used to compute the similarity score without using AReg again to speed up inference. We further propose a multi-scale GED discriminator to enhance the expressive ability of the learned representations. Extensive experiments on real-world datasets demonstrate the effectiveness, efficiency and transferability of our approach.

## Reconsidering Deep Ensembles

- Taiga Abe · Estefany Kelly Buchanan · Geoff Pleiss · Richard Zemel · John Cunningham
- abstract@[open-review](#): Ensembling neural networks is an effective way to increase accuracy, and can often match the performance of individual larger models. This observation poses a natural question: given the choice between a deep ensemble and a single neural network with similar accuracy, is one preferable over the other? Recent work suggests that deep ensembles may offer distinct benefits beyond predictive power: namely, uncertainty quantification and robustness to dataset shift. In this work, we demonstrate limitations to these purported benefits, and show that a single (but larger) neural network can replicate these qualities. First, we show that ensemble diversity, by any metric, does not meaningfully contribute to an ensemble's ability to detect out-of-distribution (OOD) data, and that one can estimate ensemble diversity by measuring the relative improvement of a single larger model. Second, we show that the OOD performance afforded by ensembles is strongly determined by their in-distribution (InD) performance, and - in this sense - is not indicative of any "effective robustness." While deep ensembles are a practical way to achieve improvements to predictive power, uncertainty quantification, and robustness, our results show that these improvements can be replicated by a (larger) single model.

## Refining Low-Resource Unsupervised Translation by Language Disentanglement of Multilingual Translation Model

- Xuan-Phi Nguyen · Shafiq Joty · Kui Wu · Ai Ti Aw
- abstract@[open-review](#): Numerous recent work on unsupervised machine translation (UMT) implies that competent unsupervised translations of low-resource and unrelated languages, such as Nepali or Sinhala, are only possible if the model is trained in a massive multilingual environment, where these low-resource languages are mixed with high-resource counterparts. Nonetheless, while the high-resource languages greatly help kick-start the target low-resource translation tasks, the language discrepancy between them may hinder their further improvement. In this work, we propose a simple refinement procedure to separate languages from a pre-trained multilingual UMT model for it to focus on only the target low-resource task. Our method achieves the state of the art in the fully unsupervised translation tasks of English to Nepali, Sinhala, Gujarati, Latvian, Estonian and Kazakh, with BLEU score gains of 3.5, 3.5, 3.3, 4.1, 4.2, and 3.3, respectively. Our codebase is available at [https://github.com/nxphi47/refineunsupmultilingual\\_mt](https://github.com/nxphi47/refineunsupmultilingual_mt)

## Nonstationary Dual Averaging and Online Fair Allocation

- Luofeng Liao · Yuan Gao · Christian Kroer
- abstract@[open-review](#): We consider the problem of fairly allocating sequentially arriving items to a set of individuals. For this problem, the recently-introduced PACE algorithm leverages the dual averaging algorithm to approximate competitive equilibria and thus generate online fair allocations. PACE is simple, distributed, and parameter-free, making it appealing for practical use in large-scale systems. However, current performance guarantees for PACE require i.i.d. item arrivals. Since real-world data is rarely i.i.d., or even stationary, we study the performance of PACE on nonstationary data. We start by developing new convergence results for the general dual averaging algorithm under three nonstationary input models: adversarially-corrupted stochastic input, ergodic input, and block-independent (including periodic) input. Our results show convergence of dual averaging up to errors caused by nonstationarity of the data, and recover the classical bounds when the input data is i.i.d. Using these results, we show that the PACE algorithm for online fair allocation simultaneously achieves ``best of many worlds'' guarantees against any of these nonstationary input models as well as against i.i.d. input. Finally, numerical experiments show strong empirical performance of PACE against nonstationary inputs.

## Old can be Gold: Better Gradient Flow can make Vanilla-GCNs Great Again

- AJAY JAISWAL · Peihao Wang · Tianlong Chen · Justin Rousseau · Ying Ding · Zhangyang Wang
- abstract@[open-review](#): Despite the enormous success of Graph Convolutional Networks (GCNs) in modeling graph-structured data, most of the current GCNs are shallow due to the notoriously challenging problems of over-smoothing and information squashing along with conventional difficulty caused by vanishing gradients and over-fitting. Previous works have been primarily focused on the study of the over-smoothing and over-squashing phenomena in training deep GCNs. Surprisingly, in comparison with CNNs/RNNs, very limited attention has been given towards understanding how healthy gradient flow can benefit the trainability of deep GCNs. In this paper, firstly, we provide a new perspective of gradient flow to understand the substandard performance of deep GCNs and hypothesize that by facilitating healthy gradient flow, we can significantly improve their trainability, as well as achieve SOTA level performance from vanilla-GCNs \cite{Kipf2017SemiSupervisedCW}. Next, we argue that blindly adopting the Glorot initialization for GCNs is not optimal, and derive a \textbf{topology-aware isometric initialization} scheme for vanilla-GCNs based on the principles of isometry. Additionally, contrary to the ad-hoc addition of skip-connections, we propose to use Dirichlet Energy for the dynamic rewiring of vanilla-GCNs with skip-connections. Our dynamic rewiring method uses the expressiveness of feature embedding learned by each layer during training to introduce skip-connections on-demand basis. We provide extensive empirical evidence across multiple datasets that our methods improve gradient flow in deep vanilla-GCNs and significantly boost their performance to comfortably compete and outperform many fancy state-of-the-art methods. Codes will be available in supplementary.

## Aligning human and machine vision

- Thomas FEL · Ivan F Rodriguez Rodriguez · Drew A Linsley · Thomas Serre
- abstract@[open-review](#): The many successes of deep neural networks (DNNs) over the past decade have largely been driven by computational scale rather than insights from biological intelligence. Here, we explore if these trends have also carried concomitant improvements in explaining visual strategies underlying human object recognition. We do this by comparing two related but distinct properties of visual strategies in humans and DNNs: where they believe important visual features are in images and how they use those features to categorize objects. Across 85 different DNNs and three independent datasets measuring human visual strategies on ImageNet, we find a trade-off between DNN top-1 categorization accuracy and their alignment with humans. State-of-the-art DNNs are progressively becoming less aligned with humans. We rectify this growing issue by introducing the neural harmonizer: a general-purpose training routine that aligns DNN and human visual strategies while improving object classification performance. Our work represents the first systematic demonstration that the scaling laws that are guiding DNN developments today have also produced worse models of human vision. We release our code and data at <https://tinyurl.com/metapred> to help the field build more human-like DNNs.

## Policy Optimization for Markov Games: Unified Framework and Faster Convergence

- Runyu Zhang · Qinghua Liu · Huan Wang · Caiming Xiong · Na Li · Yu Bai
- abstract@[open-review](#): This paper studies policy optimization algorithms for multi-agent reinforcement learning. We begin by proposing an algorithm framework for two-player zero-sum Markov Games in the full-information setting, where each iteration consists of a policy update step at each state using a certain matrix game algorithm, and a value update step with a certain learning rate. This framework unifies many existing and new policy optimization algorithms. We show that the \emph{state-wise average policy} of this algorithm converges to an approximate Nash equilibrium (NE) of the game, as long as the matrix game algorithms achieve low weighted regret at each state, with respect to weights determined by the speed of the value updates. Next, we show that this framework instantiated with the Optimistic Follow-The-Regularized-Leader (OFTL) algorithm at each state (and smooth value updates) can find an  $\mathcal{\tilde{O}}(T^{1/2})$  approximate NE in  $T$  iterations, which improves over the current best  $\mathcal{\tilde{O}}(T^{1/3})$  rate of symmetric policy optimization type algorithms. We also extend this algorithm to multi-player general-sum Markov Games and show an  $\mathcal{\tilde{O}}(T^{1/4})$  convergence rate to Coarse Correlated Equilibria (CCE). Finally, we provide a numerical example to verify our theory and investigate the importance of smooth value updates, and find that using ``eager'' value updates instead (equivalent to the independent natural policy gradient algorithm) may significantly slow down the convergence, even on a simple game with  $H=2$  layers.

## [DHRL: A Graph-Based Approach for Long-Horizon and Sparse Hierarchical Reinforcement Learning](#)

- Seungjae Lee · Jigang Kim · Inkyu Jang · H. Jin Kim
- abstract@[open-review](#): Hierarchical Reinforcement Learning (HRL) has made notable progress in complex control tasks by leveraging temporal abstraction. However, previous HRL algorithms often suffer from serious data inefficiency as environments get large. The extended components, i.e., goal space and length of episodes, impose a burden on either one or both high-level and low-level policies since both levels share the total horizon of the episode. In this paper, we present a method of Decoupling Horizons Using a Graph in Hierarchical Reinforcement Learning (DHRL) which can alleviate this problem by decoupling the horizons of high-level and low-level policies and bridging the gap between the length of both horizons using a graph. DHRL provides a freely stretchable high-level action interval, which facilitates longer temporal abstraction and faster training in complex tasks. Our method outperforms state-of-the-art HRL algorithms in typical HRL environments. Moreover, DHRL achieves long and complex locomotion and manipulation tasks.

## [Systematic improvement of neural network quantum states using Lanczos](#)

- Hongwei Chen · Douglas Hendry · Phillip Weinberg · Adrian Feiguin
- abstract@[open-review](#): The quantum many-body problem lies at the center of the most important open challenges in condensed matter, quantum chemistry, atomic, nuclear, and high-energy physics. While quantum Monte Carlo, when applicable, remains the most powerful numerical technique capable of treating dozens or hundreds of degrees of freedom with high accuracy, it is restricted to models that are not afflicted by the infamous sign problem. A powerful alternative that has emerged in recent years is the use of neural networks as variational estimators for quantum states. In this work, we propose a symmetry-projected variational solution in the form of linear combinations of simple restricted Boltzmann machines. This construction allows one to explore states outside of the original variational manifold and increase the representation power with moderate computational effort. Besides allowing one to restore spatial symmetries, an expansion in terms of Krylov states using a Lanczos recursion offers a solution that can further improve the quantum state accuracy. We illustrate these ideas with an application to the Heisenberg  $J_1-J_2$  model on the square lattice, a paradigmatic problem under debate in condensed matter physics, and achieve state-of-the-art accuracy in the representation of the ground state.

## [List-Decodable Sparse Mean Estimation via Difference-of-Pairs Filtering](#)

- Ilias Diakonikolas · Daniel Kane · Sushrut Karmalkar · Ankit Pensia · Thanasis Pittas
- abstract@[open-review](#): We study the problem of list-decodable sparse mean estimation. Specifically, for a parameter  $\alpha \in (0, 1/2)$ , we are given  $m$  points in  $\mathbb{R}^n$ ,  $\lfloor \alpha m \rfloor$  of which are i.i.d. samples from a distribution  $D$  with unknown  $k$ -sparse mean  $\mu$ . No assumptions are made on the remaining points, which form the majority of the dataset. The goal is to return a small list of candidates containing a vector  $\hat{\mu}$  such that  $\|\hat{\mu} - \mu\|_2$  is small. Prior work had studied the problem of list-decodable mean estimation in the dense setting. In this work, we develop a novel, conceptually simpler technique for list-decodable mean estimation. As the main application of our approach, we provide the first sample and computationally efficient algorithm for list-decodable sparse mean estimation. In particular, for distributions with ``certifiably bounded''  $t$ -th moments in  $k$ -sparse directions and sufficiently light tails, our algorithm achieves error of  $(1/\alpha)^{O(1/t)}$  with sample complexity  $m = (k \log(n))^{O(t)} \alpha$  and running time  $\mathcal{O}(mn^t)$ . For the special case of Gaussian inliers, our algorithm achieves the optimal error guarantee  $\Theta(\sqrt{\log(1/\alpha)})$  with quasi-polynomial complexity. We complement our upper bounds with nearly-matching statistical query and low-degree polynomial testing lower bounds.

## [No-regret learning in games with noisy feedback: Faster rates and adaptivity via learning rate separation](#)

- Yu-Guan Hsieh · Kimon Antonakopoulos · Volkan Cevher · Panayotis Mertikopoulos
- abstract@[open-review](#): We examine the problem of regret minimization when the learner is involved in a continuous game with other optimizing agents: in this case, if all players follow a no-regret algorithm, it is possible to achieve significantly lower regret relative to fully adversarial environments. We study this problem in the context of variationally stable games (a class of continuous games which includes all convex-concave and monotone games), and when the players only have access to noisy estimates of their individual payoff gradients. If the noise is additive, the game-theoretic and purely adversarial settings enjoy similar regret guarantees; however, if the noise is multiplicative, we show that the learners can, in fact, achieve constant regret. We achieve this faster rate via an optimistic gradient scheme with learning rate separation \textendash that is, the method's extrapolation and update steps are tuned to different schedules, depending on the noise profile. Subsequently, to eliminate the need for delicate hyperparameter tuning, we propose a fully adaptive method that smoothly interpolates between worst- and best-case regret guarantees.

## [Deliberated Domain Bridging for Domain Adaptive Semantic Segmentation](#)

- Lin Chen · Zhixiang Wei · Xin Jin · Huaian Chen · Miao Zheng · Kai Chen · Yi Jin
- abstract@[open-review](#): In unsupervised domain adaptation (UDA), directly adapting from the source to the target domain usually suffers significant discrepancies and leads to insufficient alignment. Thus, many UDA works attempt to vanish the domain gap gradually and softly via various intermediate spaces, dubbed domain bridging (DB). However, for dense prediction tasks such as domain adaptive semantic segmentation (DASS), existing solutions have mostly relied on rough style transfer and how to elegantly bridge domains is still under-explored. In this work, we resort to data mixing to establish a deliberated domain bridging (DDB) for DASS, through which the joint distributions of source and target domains are aligned and interacted with each in the intermediate space. At the heart of DDB lies a dual-path domain bridging step for generating two intermediate domains using the coarse-wise and the fine-wise data mixing techniques, alongside a cross-path knowledge distillation step for taking two complementary models trained on generated intermediate samples as teachers to develop a superior student in a multi-teacher distillation manner. These two optimization steps work in an alternating way and reinforce each other to give rise to DDB with strong adaptation power. Extensive experiments on adaptive segmentation tasks with different settings demonstrate that our DDB significantly outperforms state-of-the-art methods.

## [Lower Bounds on Randomly Preconditioned Lasso via Robust Sparse Designs](#)

- Jonathan Kelner · Frederic Koehler · Raghu Meka · Dhruv Rohatgi

- abstract@[open-review](#): Sparse linear regression with ill-conditioned Gaussian random covariates is widely believed to exhibit a statistical/computational gap, but there is surprisingly little formal evidence for this belief. Recent work has shown that, for certain covariance matrices, the broad class of Preconditioned Lasso programs provably cannot succeed on polylogarithmically sparse signals with a sublinear number of samples. However, this lower bound only holds against deterministic preconditioners, and in many contexts randomization is crucial to the success of preconditioners. We prove a stronger lower bound that rules out randomized preconditioners. For an appropriate covariance matrix, we construct a single signal distribution on which any invertibly-preconditioned Lasso program fails with high probability, unless it receives a linear number of samples. Surprisingly, at the heart of our lower bound is a new robustness result in compressed sensing. In particular, we study recovering a sparse signal when a few measurements can be erased adversarially. To our knowledge, this natural question has not been studied before for sparse measurements. We surprisingly show that standard sparse Bernoulli measurements are almost-optimally robust to adversarial erasures: if  $\$b\$$  measurements are erased, then all but  $\$O(b)\$$  of the coordinates of the signal are identifiable.

## [Coreset for Line Sets Clustering](#)

- Sagi Lotan Â· Ernesto Evgeniy Sanches Shayda Â· Dan Feldman
- abstract@[open-review](#): The input to the {line-sets \$k\$-median} problem is an integer  $k \geq 1$ , and a set  $\mathcal{L} = \{L_1, \dots, L_n\}$  of  $n$  sets of lines in  $\mathbb{R}^d$ . The goal is to compute a set  $C$  of  $k$  centers (points in  $\mathbb{R}^d$ ) that minimizes the sum  $\sum_{L \in \mathcal{L}} \min_{c \in C} \text{dist}(L, c)$  of Euclidean distances from each set to its closest center, where  $\text{dist}(L, c) := \min_{x \in L} \|x - c\|_2$ . An  $\epsilon$ -coreset for this problem is a weighted subset of sets in  $\mathcal{L}$  that approximates this sum up to  $(1 + \epsilon) \cdot \text{dist}(C, c)$  multiplicative factor, for every set  $C$  of  $k$  centers. We prove that every such input set  $\mathcal{L}$  has a small  $\epsilon$ -coreset, and provide the first coreset construction for this problem and its variants. The coreset consists of  $O(\log^2 n)$  weighted line-sets from  $\mathcal{L}$ , and is constructed in  $O(n \log n)$  time for every fixed  $d, k \geq 1$  and  $\epsilon \in (0, 1)$ . The main technique is based on a novel reduction to a ``fair clustering'' of colored points to colored centers. We then provide a coreset for this coloring problem, which may be of independent interest. Open source code and experiments are also provided.

## [When does dough become a bagel? Analyzing the remaining mistakes on ImageNet](#)

- Vijay Vasudevan Â· Benjamin Caine Â· Raphael Gontijo Lopes Â· Sara Fridovich-Keil Â· Rebecca Roelofs
- abstract@[open-review](#): Image classification accuracy on the ImageNet dataset has been a barometer for progress in computer vision over the last decade. Several recent papers have questioned the degree to which the benchmark remains useful to the community, yet innovations continue to contribute gains to performance, with today's largest models achieving 90%+ top-1 accuracy. To help contextualize progress on ImageNet and provide a more meaningful evaluation for today's state-of-the-art models, we manually review and categorize every remaining mistake that a few top models make in order to provide insight into the long-tail of errors on one of the most benchmarked datasets in computer vision. We focus on the multi-label subset evaluation of ImageNet, where today's best models achieve upwards of 97% top-1 accuracy. Our analysis reveals that nearly half of the supposed mistakes are not mistakes at all, and we uncover new valid multi-labels, demonstrating that, without careful review, we are significantly underestimating the performance of these models. On the other hand, we also find that today's best models still make a significant number of mistakes (40%) that are obviously wrong to human reviewers. To calibrate future progress on ImageNet, we provide an updated multi-label evaluation set, and we curate ImageNet-Major: a 68-example "major error" slice of the obvious mistakes made by today's top models -- a slice where models should achieve near perfection, but today are far from doing so.

## [First Hitting Diffusion Models](#)

- Mao Ye Â· Lemeng Wu Â· Qiang Liu
- abstract@[open-review](#): We propose a family of First Hitting Diffusion Models (FHDM), deep generative models that generate data with a diffusion process that terminates at a random first hitting time. This yields an extension of the standard fixed-time diffusion models that terminate at a pre-specified deterministic time. Although standard diffusion models are designed for continuous unconstrained data, FHDM is naturally designed to learn distributions on continuous as well as a range of discrete and structure domains. Moreover, FHDM enables instance-dependent terminate time and accelerates the diffusion process to sample higher quality data with fewer diffusion steps. Technically, we train FHDM by maximum likelihood estimation on diffusion trajectories augmented from observed data with conditional first hitting processes (i.e., bridge) derived based on Doob's  $h$ -transform, deviating from the commonly used time-reversal mechanism. We apply FHDM to generate data in various domains such as point cloud (general continuous distribution), climate and geographical events on earth (continuous distribution on the sphere), unweighted graphs (distribution of binary matrices), and segmentation maps of 2D images (high-dimensional categorical distribution). We observe considerable improvement compared with the state-of-the-art approaches in both quality and speed.

## [Nearly Optimal Algorithms for Linear Contextual Bandits with Adversarial Corruptions](#)

- Jiafan He Â· Dongruo Zhou Â· Tong Zhang Â· Quanquan Gu
- abstract@[open-review](#): We study the linear contextual bandit problem in the presence of adversarial corruption, where the reward at each round is corrupted by an adversary, and the corruption level (i.e., the sum of corruption magnitudes over the horizon) is  $C \geq 0$ . The best-known algorithms in this setting are limited in that they either are computationally inefficient or require a strong assumption on the corruption, or their regret is at least  $C$  times worse than the regret without corruption. In this paper, to overcome these limitations, we propose a new algorithm based on the principle of optimism in the face of uncertainty. At the core of our algorithm is a weighted ridge regression where the weight of each chosen action depends on its confidence up to some threshold. We show that for both known  $C$  and unknown  $C$  cases, our algorithm with proper choice of hyperparameter achieves a regret that nearly matches the lower bounds. Thus, our algorithm is nearly optimal up to logarithmic factors for both cases. Notably, our algorithm achieves the near-optimal regret for both corrupted and uncorrupted cases ( $C=0$ ) simultaneously.

## [Adaptation Accelerating Sampling-based Bayesian Inference in Attractor Neural Networks](#)

- Xingsi Dong Â· Zilong Ji Â· Tianhao Chu Â· Tiejun Huang Â· Wenhao Zhang Â· Si Wu
- abstract@[open-review](#): The brain performs probabilistic Bayesian inference to interpret the external world. The sampling-based view assumes that the brain represents the stimulus posterior distribution via samples of stochastic neuronal responses. Although the idea of sampling-based inference is appealing, it faces a critical challenge of whether stochastic sampling is fast enough to match the rapid computation of the brain. In this study, we explore how latent stimulus sampling can be accelerated in neural circuits. Specifically, we consider a canonical neural circuit model called continuous attractor neural networks (CANNs) and investigate how sampling-based inference of latent continuous variables is accelerated in CANNs. Intriguingly, we find that by including noisy adaptation in the neuronal dynamics, the CANN is able to speed up the sampling process significantly. We theoretically derive that the CANN with noisy adaptation implements the efficient sampling method called Hamiltonian dynamics with friction, where noisy adaption effectively plays the role of momentum. We theoretically analyze the sampling performances of the network and derive the condition when the acceleration has the maximum effect. Simulation results confirm our theoretical analyses. We further extend the model to coupled CANNs and demonstrate that noisy adaptation accelerates the sampling of the posterior distribution of multivariate stimuli. We hope that this study enhances our understanding of how Bayesian inference is realized in the brain.

## [Learning to Discover and Detect Objects](#)

- Vladimir Fomenko · Ismail Elezi · Deva Ramanan · Aljosa Osep · Laura Leal-Taix ©
- abstract@[open-review](#): We tackle the problem of novel class discovery, detection, and localization (NCDL). In this setting, we assume a source dataset with labels for objects of commonly observed classes. Instances of other classes need to be discovered, classified, and localized automatically based on visual similarity, without human supervision. To this end, we propose a two-stage object detection network Region-based NCDL (RNCDL), that uses a region proposal network to localize potential objects and classify them. We train our network to classify each proposal, either as one of the known classes, seen in the source dataset, or one of the extended set of novel classes with a constraint that the distribution of class assignments should follow natural long-tail distributions common in the real open-world. By training our detection network with this objective in an end-to-end manner, it learns to classify all region proposals for a large variety of classes, including those that are not part of the labeled object class vocabulary. Our experiments conducted using COCO and LVIS datasets reveal that our method is significantly more effective compared to multi-stage pipelines that rely on traditional clustering algorithms or self-supervised contrastive learning methods and operate on pre-extracted crops. Beyond that, we demonstrate the generality of our approach by applying our method to a large-scale Visual Genome dataset, where our network successfully learns to detect various semantic classes without explicit supervision.

## [Neural Circuit Architectural Priors for Embodied Control](#)

- Nikhil Bhattachariya · Anthony M Zador · Tatiana Engel
- abstract@[open-review](#): Artificial neural networks for motor control usually adopt generic architectures like fully connected MLPs. While general, these tabula rasa architectures rely on large amounts of experience to learn, are not easily transferable to new bodies, and have internal dynamics that are difficult to interpret. In nature, animals are born with highly structured connectivity in their nervous systems shaped by evolution; this innate circuitry acts synergistically with learning mechanisms to provide inductive biases that enable most animals to function well soon after birth and learn efficiently. Convolutional networks inspired by visual circuitry have encoded useful biases for vision. However, it is unknown the extent to which ANN architectures inspired by neural circuitry can yield useful biases for other AI domains. In this work, we ask what advantages biologically inspired network architecture can provide in the context of motor control. Specifically, we translate *C. elegans* locomotion circuits into an ANN model controlling a simulated Swimmer agent. On a locomotion task, our architecture achieves good initial performance and asymptotic performance comparable with MLPs, while dramatically improving data efficiency and requiring orders of magnitude fewer parameters. Our architecture is more interpretable and transfers to new body designs. An ablation analysis shows that constrained excitation/inhibition is crucial for learning, while weight initialization contributes to good initial performance. Our work demonstrates several advantages of an ANN architecture inspired by systems neuroscience and encourages future work on more complex AI motor control bodies.

## [On Batch Teaching with Sample Complexity Bounded by VCD](#)

- Farnam Mansouri · Hans Simon · Adish Singla · Sandra Zilles
- abstract@[open-review](#): In machine teaching, a concept is represented by (and inferred from) a small number of labeled examples. Various teaching models in the literature cast the interaction between teacher and learner in a way to obtain a small complexity (in terms of the number of examples required for teaching a concept) while obeying certain constraints that are meant to prevent unfair collusion between teacher and learner. In recent years, one major research goal has been to show interesting relationships between teaching complexity and the VC-dimension (VCD). So far, the only interesting relationship known from batch teaching settings is an upper bound quadratic in the VCD, on a parameter called recursive teaching dimension. The only known upper bound on teaching complexity that is linear in VCD was obtained in a model of teaching with sequences rather than batches. This paper is the first to provide an upper bound of VCD on a batch teaching complexity parameter. This parameter, called STDmin, is introduced here as a model of teaching that intuitively incorporates a notion of ``importance'' of an example for a concept. In designing the STDmin teaching model, we argue that the standard notion of collusion-freeness from the literature may be inadequate for certain applications; we hence propose three desirable properties of teaching complexity and demonstrate that they are satisfied by STDmin.

## [Partial Identification of Treatment Effects with Implicit Generative Models](#)

- Vahid Balazadeh Meresht · Vasilis Syrgkanis · Rahul Krishnan
- abstract@[open-review](#): We propose a new method for the problem of partial identification, the estimation of bounds on the treatment effects from observational data. Although studied using discrete treatment variables or in specific causal graphs (e.g., instrumental variables), partial identification has been recently explored using tools from deep generative modeling. We propose a new method for partial identification of average treatment effects (ATEs) in general causal graphs using implicit generative models comprising continuous and discrete random variables. We leverage average treatment derivatives, the partial derivatives of response functions, to prove that our algorithm converges to tight bounds on ATE. Our empirical results show that using average treatment derivatives leads to tighter and more stable bounds than methods that directly optimize the ATE when treatments are continuous. In the case of discrete treatments, our derived bounds match those from bespoke solutions for partial identification.

## [Non-Convex Bilevel Games with Critical Point Selection Maps](#)

- Michael Arbel · Julien Mairal
- abstract@[open-review](#): Bilevel optimization problems involve two nested objectives, where an upper-level objective depends on a solution to a lower-level problem. When the latter is non-convex, multiple critical points may be present, leading to an ambiguous definition of the problem. In this paper, we introduce a key ingredient for resolving this ambiguity through the concept of a selection map which allows one to choose a particular solution to the lower-level problem. Using such maps, we define a class of hierarchical games between two agents that resolve the ambiguity in bilevel problems. This new class of games requires introducing new analytical tools in Morse theory to characterize their evolution. In particular, we study the differentiability of the selection, an essential property when analyzing gradient-based algorithms for solving these games. We show that many existing algorithms for bilevel optimization, such as unrolled optimization, solve these games up to approximation errors due to finite computational power. Our analysis allows introducing a simple correction to these algorithms for removing the errors.

## [Domain Generalization without Excess Empirical Risk](#)

- Ozan Sener · Vladlen Koltun
- abstract@[open-review](#): Given data from diverse sets of distinct distributions, domain generalization aims to learn models that generalize to unseen distributions. A common approach is designing a data-driven surrogate penalty to capture generalization and minimize the empirical risk jointly with the penalty. We argue that a significant failure mode of this recipe is an excess risk due to an erroneous penalty or hardness in joint optimization. We present an approach that eliminates this problem. Instead of jointly minimizing empirical risk with the penalty, we minimize the penalty under the constraint of optimality of the empirical risk. This change guarantees that the domain generalization penalty cannot impair optimization of the empirical risk, i.e., in-distribution performance. To solve the proposed optimization problem, we demonstrate an exciting connection to rate-distortion theory and utilize its tools to design an efficient method. Our approach can be applied to any penalty-based domain generalization method, and we demonstrate its effectiveness by applying it to three exemplar methods from the literature, showing significant improvements.

## [STaR: Bootstrapping Reasoning With Reasoning](#)

- Eric Zelikman · Yuhuai Wu · Jesse Mu · Noah Goodman

- abstract@[open-review](#): Generating step-by-step "chain-of-thought" rationales improves language model performance on complex reasoning tasks like mathematics or commonsense question-answering. However, inducing language model rationale generation currently requires either constructing massive rationale datasets or sacrificing accuracy by using only few-shot inference. We propose a technique to iteratively leverage a small number of rationale examples and a large dataset without rationales, to bootstrap the ability to perform successively more complex reasoning. This technique, the "Self-Taught Reasoner" (STaR), relies on a simple loop: generate rationales to answer many questions, prompted with a few rationale examples; if the generated answers are wrong, try again to generate a rationale given the correct answer; fine-tune on all the rationales that ultimately yielded correct answers; repeat. We show that STaR significantly improves performance on multiple datasets compared to a model fine-tuned to directly predict final answers, and performs comparably to fine-tuning a 30\$times\$ larger state-of-the-art language model on CommonsenseQA. Thus, STaR lets a model improve itself by learning from its own generated reasoning.

## [Continual Learning In Environments With Polynomial Mixing Times](#)

- Matthew Riemer Â· Sharath Chandra Raparthy Â· Ignacio Cases Â· Gopesh Subbaraj Â· Maximilian Puelma Touzel Â· Irina Rish
- abstract@[open-review](#): The mixing time of the Markov chain induced by a policy limits performance in real-world continual learning scenarios. Yet, the effect of mixing times on learning in continual reinforcement learning (RL) remains underexplored. In this paper, we characterize problems that are of long-term interest to the development of continual RL, which we call scalable MDPs, through the lens of mixing times. In particular, we theoretically establish that scalable MDPs have mixing times that scale polynomially with the size of the problem. We go on to demonstrate that polynomial mixing times present significant difficulties for existing approaches that suffer from myopic bias and stale bootstrapped estimates. To validate the proposed theory, we study the empirical scaling behavior of mixing times with respect to the number of tasks and task switching frequency for pretrained high performing policies on seven Atari games. Our analysis demonstrates both that polynomial mixing times do emerge in practice and how their existence may lead to unstable learning behavior like catastrophic forgetting in continual learning settings.

## [Learning from Label Proportions by Learning with Label Noise](#)

- Jianxin Zhang Â· Yutong Wang Â· Clay Scott
- abstract@[open-review](#): Learning from label proportions (LLP) is a weakly supervised classification problem where data points are grouped into bags, and the label proportions within each bag are observed instead of the instance-level labels. The task is to learn a classifier to predict the labels of future individual instances. Prior work on LLP for multi-class data has yet to develop a theoretically grounded algorithm. In this work, we propose an approach to LLP based on a reduction to learning with label noise, using the forward correction (FC) loss of \textcite{Patrini2017MakingDN}. We establish an excess risk bound and generalization error analysis for our approach, while also extending the theory of the FC loss which may be of independent interest. Our approach demonstrates improved empirical performance in deep learning scenarios across multiple datasets and architectures, compared to the leading methods.

## [Self-Similarity Priors: Neural Collages as Differentiable Fractal Representations](#)

- Michael Poli Â· Winnie Xu Â· Stefano Massaroli Â· Chenlin Meng Â· Kuno Kim Â· Stefano Ermon
- abstract@[open-review](#): Many patterns in nature exhibit self-similarity: they can be compactly described via self-referential transformations. Said patterns commonly appear in natural and artificial objects, such as molecules, shorelines, galaxies, and even images. In this work, we investigate the role of learning in the automated discovery of self-similarity and in its utilization for downstream tasks. To this end, we design a novel class of implicit operators, Neural Collages, which (1) represent data as the parameters of a self-referential, structured transformation, and (2) employ hypernetworks to amortize the cost of finding these parameters to a single forward pass. We investigate how to leverage the representations produced by Neural Collages in various tasks, including data compression and generation. Neural Collage image compressors are orders of magnitude faster than other self-similarity-based algorithms during encoding and offer compression rates competitive with implicit methods. Finally, we showcase applications of Neural Collages for fractal art and as deep generative models.

## [Optimal Efficiency-Envy Trade-Off via Optimal Transport](#)

- Steven Yin Â· Christian Kroer
- abstract@[open-review](#): We consider the problem of allocating a distribution of items to \$n\$ recipients where each recipient has to be allocated a fixed, pre-specified fraction of all items, while ensuring that each recipient does not experience too much envy. We show that this problem can be formulated as a variant of the semi-discrete optimal transport (OT) problem, whose solution structure in this case has a concise representation and a simple geometric interpretation. Unlike existing literature that treats envy-freeness as a hard constraint, our formulation allows us to \emph{optimally} trade off efficiency and envy continuously. Additionally, we study the statistical properties of the space of our OT based allocation policies by showing a polynomial bound on the number of samples needed to approximate the optimal solution from samples. Our approach is suitable for large-scale fair allocation problems such as the blood donation matching problem, and we show numerically that it performs well on a prior realistic data simulator.

## [On the relationship between variational inference and auto-associative memory](#)

- Louis Annabi Â· Alexandre Pitti Â· Mathias Quoy
- abstract@[open-review](#): In this article, we propose a variational inference formulation of memory retrieval in auto-associative memories, allowing us to combine memory retrieval with perceptual inference into the same mathematical framework. In this formulation, the prior probability distribution onto representations is made memory dependent, thus pulling the inference process towards stored representations. We then study how different neural network approaches to variational inference can be applied in this framework. We compare methods relying on amortized inference such as Variational Autoencoders and methods relying on iterative inference such as Predictive Coding and suggest combining both approaches to design new auto-associative memory models. We evaluate the obtained algorithms on the CIFAR10 and CLEVR image datasets and compare them with other associative memory models such as Hopfield Networks, End-to-End Memory Networks and Neural Turing Machines.

## [Bounded-Regret MPC via Perturbation Analysis: Prediction Error, Constraints, and Nonlinearity](#)

- Yiheng Lin Â· Yang Hu Â· Guannan Qu Â· Tongxin Li Â· Adam Wierman
- abstract@[open-review](#): We study Model Predictive Control (MPC) and propose a general analysis pipeline to bound its dynamic regret. The pipeline first requires deriving a perturbation bound for a finite-time optimal control problem. Then, the perturbation bound is used to bound the per-step error of MPC, which leads to a bound on the dynamic regret. Thus, our pipeline reduces the study of MPC to the well-studied problem of perturbation analysis, enabling the derivation of regret bounds of MPC under a variety of settings. To demonstrate the power of our pipeline, we use it to generalize existing regret bounds on MPC in linear time-varying (LTV) systems to incorporate prediction errors on costs, dynamics, and disturbances. Further, our pipeline leads to regret bounds on MPC in systems with nonlinear dynamics and constraints.

## [Reduced Representation of Deformation Fields for Effective Non-rigid Shape Matching](#)

- Ramana Subramanyam Sundararaman Â· Riccardo Marin Â· Emanuele RodolÃ Â· Maks Ovsjanikov

- abstract@[open-review](#): In this work we present a novel approach for computing correspondences between non-rigid objects, by exploiting a reduced representation of deformation fields. Different from existing works that represent deformation fields by training a general-purpose neural network, we advocate for an approximation based on mesh-free methods. By letting the network learn deformation parameters at a sparse set of positions in space (nodes), we reconstruct the continuous deformation field in a closed-form with guaranteed smoothness. With this reduction in degrees of freedom, we show significant improvement in terms of data-efficiency and enable limited supervision. Furthermore, our approximation provides direct access to first-order derivatives of deformation fields, which facilitates enforcing desirable regularization effectively. Our resulting model has a high expressive power and is able to capture complex deformations. We illustrate its effectiveness through state-of-the-art results across multiple deformable shape matching benchmarks.

## [A Reparametrization-Invariant Sharpness Measure Based on Information Geometry](#)

- Cheongjae Jang · Sungyoon Lee · Yung-Kyun Noh · Frank Park
- abstract@[open-review](#): It has been observed that the generalization performance of neural networks correlates with the sharpness of their loss landscape. Dinh et al (2017) have observed that existing formulations of sharpness measures fail to be invariant with respect to scaling and reparametrization. While some scale-invariant measures have recently been proposed, reparametrization-invariant measures are still lacking, as are any theoretical insights into generalization performance. Based on an information geometric analysis of the neural network parameter space, in this paper we propose a reparametrization-invariant sharpness measure that captures the change in loss with respect to changes in the probability distribution modeled by neural networks, rather than with respect to changes in the parameter values. We reveal some theoretical connections of our measure to generalization performance. In particular, experiments confirm that using our measure as a regularizer in neural network training significantly improves performance.

## [Stability Analysis and Generalization Bounds of Adversarial Training](#)

- Jiancong Xiao · Yanbo Fan · Ruoyu Sun · Jue Wang · Zhi-Quan Luo
- abstract@[open-review](#): In adversarial machine learning, deep neural networks can fit the adversarial examples on the training dataset but have poor generalization ability on the test set. This phenomenon is called robust overfitting, and it can be observed when adversarially training neural nets on common datasets, including SVHN, CIFAR-10, CIFAR-100, and ImageNet. In this paper, we study the robust overfitting issue of adversarial training by using tools from uniform stability. One major challenge is that the outer function (as a maximization of the inner function) is nonsmooth, so the standard technique (e.g., [19]) cannot be applied. Our approach is to consider  $\eta$ -approximate smoothness: we show that the outer function satisfies this modified smoothness assumption with  $\eta$  being a constant related to the adversarial perturbation. Based on this, we derive stability-based generalization bounds for stochastic gradient descent (SGD) on the general class of  $\eta$ -approximate smooth functions, which covers the adversarial loss. Our results provide a different understanding of robust overfitting from the perspective of uniform stability. Additionally, we show that a few popular techniques for adversarial training (e.g., early stopping, cyclic learning rate, and stochastic weight averaging) are stability-promoting in theory.

## [Mining Multi-Label Samples from Single Positive Labels](#)

- Youngin Cho · Daejin Kim · MOHAMMAD AZAM KHAN · Jaegul Choo
- abstract@[open-review](#): Conditional generative adversarial networks (cGANs) have shown superior results in class-conditional generation tasks. In order to simultaneously control multiple conditions, cGANs require multi-label training datasets, where multiple labels can be assigned to each data instance. Nevertheless, the tremendous annotation cost limits the accessibility of multi-label datasets in the real-world scenarios. Hence, we explore the practical setting called single positive setting, where each data instance is annotated by only one positive label with no explicit negative labels. To generate multi-label data in the single positive setting, we propose a novel sampling approach called single-to-multi-label (S2M) sampling, based on the Markov chain Monte Carlo method. As a widely applicable  $\epsilon$ -add-on method, our proposed S2M sampling enables existing unconditional and conditional GANs to draw high-quality multi-label data with a minimal annotation cost. Extensive experiments on real image datasets verify the effectiveness and correctness of our method, even when compared to a model trained with fully annotated datasets.

## [Pruning Neural Networks via Coresets and Convex Geometry: Towards No Assumptions](#)

- Murad Tukan · Loay Mualem · Alaa Maalouf
- abstract@[open-review](#): Pruning is one of the predominant approaches for compressing deep neural networks (DNNs). Lately, coresets (provable data summarizations) were leveraged for pruning DNNs, adding the advantage of theoretical guarantees on the trade-off between the compression rate and the approximation error. However, coresets in this domain were either data dependant or generated under restrictive assumptions on both the model's weights and inputs. In real-world scenarios, such assumptions are rarely satisfied, limiting the applicability of coresets. To this end, we suggest a novel and robust framework for computing such coresets under mild assumptions on the model's weights and without any assumption on the training data. The idea is to compute the importance of each neuron in each layer with respect to the output of the following layer. This is achieved by an elegant combination of Lowner ellipsoid and Caratheodory theorem. Our method is simultaneously data-independent, applicable to various networks and datasets (due to the simplified assumptions), and theoretically supported. Experimental results show that our method outperforms existing coreset based neural pruning approaches across a wide range of networks and datasets. For example, our method achieved a 62% compression rate on ResNet50 on ImageNet with 1.09% drop in accuracy.

## [Uncoupled Learning Dynamics with \$O\(\log T\)\$ Swap Regret in Multiplayer Games](#)

- Ioannis Anagnostides · Gabriele Farina · Christian Kroer · Chung-Wei Lee · Haipeng Luo · Tuomas Sandholm
- abstract@[open-review](#): In this paper we establish efficient and uncoupled learning dynamics so that, when employed by all players in a general-sum multiplayer game, the swap regret of each player after  $T$  repetitions of the game is bounded by  $O(\log T)$ , improving over the prior best bounds of  $O(\log^4(T))$ . At the same time, we guarantee optimal  $O(\sqrt{T})$  swap regret in the adversarial regime as well. To obtain these results, our primary contribution is to show that when all players follow our dynamics with a time-invariant learning rate, the second-order path lengths of the dynamics up to time  $T$  are bounded by  $O(\log T)$ , a fundamental property which could have further implications beyond near-optimally bounding the (swap) regret. Our proposed learning dynamics combine in a novel way optimistic regularized learning with the use of self-concordant barriers. Further, our analysis is remarkably simple, bypassing the cumbersome framework of higher-order smoothness recently developed by Daskalakis, Fishelson, and Golowich (NeurIPS'21).

## [Decomposable Non-Smooth Convex Optimization with Nearly-Linear Gradient Oracle Complexity](#)

- Sally Dong · Haotian Jiang · Yin Tat Lee · Swati Padmanabhan · Guanghao Ye
- abstract@[open-review](#): Many fundamental problems in machine learning can be formulated by the convex program  $\min_{\theta} \sum_{i=1}^n f_i(\theta)$ , where each  $f_i$  is a convex, Lipschitz function supported on a subset of  $d_i$  coordinates of  $\theta$ . One common approach to this problem, exemplified by stochastic gradient descent, involves sampling one  $f_i$  term at every iteration to make progress. This approach crucially relies on a notion of uniformity across the  $f_i$ 's, formally captured by their condition number. In this work, we give an algorithm that minimizes the above convex formulation to  $\tilde{\epsilon}$ -accuracy in  $\tilde{O}(\sum_{i=1}^n d_i \log(1/\epsilon))$  gradient computations, with no assumptions on the condition number. The previous best algorithm independent of the condition number is the standard cutting plane method, which requires  $O(nd \log(1/\epsilon))$  gradient computations. As a corollary, we improve upon the evaluation oracle complexity for decomposable submodular

minimization by [Axiotis, Karczmarz, Mukherjee, Sankowski and Vladu, ICML 2021]. Our main technical contribution is an adaptive procedure to select an  $\$f_i$  term at every iteration via a novel combination of cutting-plane and interior-point methods.

## [Robust Imitation of a Few Demonstrations with a Backwards Model](#)

- Jung Yeon Park · Lawson Wong
- abstract@[open-review](#): Behavior cloning of expert demonstrations can speed up learning optimal policies in a more sample-efficient way over reinforcement learning. However, the policy cannot extrapolate well to unseen states outside of the demonstration data, creating covariate shift (agent drifting away from demonstrations) and compounding errors. In this work, we tackle this issue by extending the region of attraction around the demonstrations so that the agent can learn how to get back onto the demonstrated trajectories if it veers off-course. We train a generative backward dynamics model and generate short imagined trajectories from states in the demonstrations. By imitating both demonstrations and these model rollouts, the agent learns both the demonstrated paths and how to get back on to these paths. With optimal or near-optimal demonstrations, the learned policy will be both optimal and robust to deviations, with a wider region of attraction. On continuous control domains, we evaluate the robustness when starting from different initial states unseen in the demonstration data. While both our method and other imitation learning baselines can successfully solve the tasks for initial states in the training distribution, our method exhibits considerably more robustness to different initial states.

## [Distributed Learning of Conditional Quantiles in the Reproducing Kernel Hilbert Space](#)

- Heng Lian
- abstract@[open-review](#): We study distributed learning of nonparametric conditional quantiles with Tikhonov regularization in a reproducing kernel Hilbert space (RKHS). Although distributed parametric quantile regression has been investigated in several existing works, the current nonparametric quantile setting poses different challenges and is still unexplored. The difficulty lies in the illusive explicit bias-variance decomposition in the quantile RKHS setting as in the regularized least squares regression. For the simple divide-and-conquer approach that partitions the data set into multiple parts and then takes an arithmetic average of the individual outputs, we establish the risk bounds using a novel second-order empirical process for quantile risk.

## [Wasserstein Iterative Networks for Barycenter Estimation](#)

- Alexander Korotin · Vage Egiazarian · Lingxiao Li · Evgeny Burnaev
- abstract@[open-review](#): Wasserstein barycenters have become popular due to their ability to represent the average of probability measures in a geometrically meaningful way. In this paper, we present an algorithm to approximate the Wasserstein-2 barycenters of continuous measures via a generative model. Previous approaches rely on regularization (entropic/quadratic) which introduces bias or on input convex neural networks which are not expressive enough for large-scale tasks. In contrast, our algorithm does not introduce bias and allows using arbitrary neural networks. In addition, based on the celebrity faces dataset, we construct Ave\_celeba! dataset which can be used for quantitative evaluation of barycenter algorithms by using standard metrics of generative models such as FID.

## [Transformers from an Optimization Perspective](#)

- Yongyi Yang · zengfeng Huang · David P Wipf
- abstract@[open-review](#): Deep learning models such as the Transformer are often constructed by heuristics and experience. To provide a complementary foundation, in this work we study the following problem: Is it possible to find an energy function underlying the Transformer model, such that descent steps along this energy correspond with the Transformer forward pass? By finding such a function, we can reinterpret Transformers as the unfolding of an interpretable optimization process. This unfolding perspective has been frequently adopted in the past to elucidate more straightforward deep models such as MLPs and CNNs; however, it has thus far remained elusive obtaining a similar equivalence for more complex models with self-attention mechanisms like the Transformer. To this end, we first outline several major obstacles before providing companion techniques to at least partially address them, demonstrating for the first time a close association between energy function minimization and deep layers with self-attention. This interpretation contributes to our intuition and understanding of Transformers, while potentially laying the ground-work for new model designs.

## [Unsupervised Skill Discovery via Recurrent Skill Training](#)

- Zheyuan Jiang · Jingyue Gao · Jianyu Chen
- abstract@[open-review](#): Being able to discover diverse useful skills without external reward functions is beneficial in reinforcement learning research. Previous unsupervised skill discovery approaches mainly train different skills in parallel. Although impressive results have been provided, we found that parallel training procedure can sometimes block exploration when the state visited by different skills overlap, which leads to poor state coverage and restricts the diversity of learned skills. In this paper, we take a deeper look into this phenomenon and propose a novel framework to address this issue, which we call Recurrent Skill Training (ReST). Instead of training all the skills in parallel, ReST trains different skills one after another recurrently, along with a state coverage based intrinsic reward. We conduct experiments on a number of challenging 2D navigation environments and robotic locomotion environments. Evaluation results show that our proposed approach outperforms previous parallel training approaches in terms of state coverage and skill diversity. Videos of the discovered skills are available at <https://sites.google.com/view/neurips22-rest>.

## [Para-CFlows: \$C^k\$ -universal diffeomorphism approximators as superior neural surrogates](#)

- Junlong Lyu · Zhitang Chen · Chang Feng · Wenjing Cun · Shengyu Zhu · Yanhui Geng · ZHIJIE XU · Chen Yongwei
- abstract@[open-review](#): Invertible neural networks based on Coupling Flows (CFlows) have various applications such as image synthesis and data compression. The approximation universality for CFlows is of paramount importance to ensure the model expressiveness. In this paper, we prove that CFlows can approximate any diffeomorphism in  $C^k$ -norm if its layers can approximate certain single-coordinate transforms. Specifically, we derive that a composition of affine coupling layers and invertible linear transforms achieves this universality. Furthermore, in parametric cases where the diffeomorphism depends on some extra parameters, we prove the corresponding approximation theorems for parametric coupling flows named Para-CFlows. In practice, we apply Para-CFlows as a neural surrogate model in contextual Bayesian optimization tasks, to demonstrate its superiority over other neural surrogate models in terms of optimization performance and gradient approximations.

## [A Data-Augmentation Is Worth A Thousand Samples](#)

- Randall Balestriero · Ishan Misra · Yann LeCun
- abstract@[open-review](#): Data-Augmentation (DA) is known to improve performance across tasks and datasets. We propose a method to theoretically analyze the effect of DA and study questions such as: how many augmented samples are needed to correctly estimate the information encoded by that DA? How does the augmentation policy impact the final parameters of a model? We derive several quantities in close-form, such as the expectation and variance of an image, loss, and model's output under a given DA distribution. Up to our knowledge, we obtain the first explicit regularizer that corresponds to using DA during training for non-trivial transformations such as affine transformations, color jittering, or Gaussian blur. Those derivations open new avenues to quantify the benefits and limitations of DA. For example, we show that common DAs require tens of thousands of samples for the loss at hand to be correctly estimated and for the model training to converge. We show that for a training loss to be stable under DA sampling, the model's

saliency map (gradient of the loss with respect to the model's input) must align with the smallest eigenvector of the sample variance under the considered DA augmentation, hinting at a possible explanation on why models tend to shift their focus from edges to textures.

## [A Closer Look at the Adversarial Robustness of Deep Equilibrium Models](#)

- Zonghan Yang · Tianyu Pang · Yang Liu
- abstract@[open-review](#): Deep equilibrium models (DEQs) refrain from the traditional layer-stacking paradigm and turn to find the fixed point of a single layer. DEQs have achieved promising performance on different applications with featured memory efficiency. At the same time, the adversarial vulnerability of DEQs raises concerns. Several works propose to certify robustness for monotone DEQs. However, limited efforts are devoted to studying empirical robustness for general DEQs. To this end, we observe that an adversarially trained DEQ requires more forward steps to arrive at the equilibrium state, or even violates its fixed-point structure. Besides, the forward and backward tracks of DEQs are misaligned due to the black-box solvers. These facts cause gradient obfuscation when applying the ready-made attacks to evaluate or adversarially train DEQs. Given this, we develop approaches to estimate the intermediate gradients of DEQs and integrate them into the attacking pipelines. Our approaches facilitate fully white-box evaluations and lead to effective adversarial defense for DEQs. Extensive experiments on CIFAR-10 validate the adversarial robustness of DEQs competitive with deep networks of similar sizes.

## [On the Effective Number of Linear Regions in Shallow Univariate ReLU Networks: Convergence Guarantees and Implicit Bias](#)

- Itay Safran · Gal Vardi · Jason Lee
- abstract@[open-review](#): We study the dynamics and implicit bias of gradient flow (GF) on univariate ReLU neural networks with a single hidden layer in a binary classification setting. We show that when the labels are determined by the sign of a target network with  $r$  neurons, with high probability over the initialization of the network and the sampling of the dataset, GF converges in direction (suitably defined) to a network achieving perfect training accuracy and having at most  $\mathcal{O}(r)$  linear regions, implying a generalization bound. Our result may already hold for mild over-parameterization, where the width is  $\tilde{\mathcal{O}}(r)$  and independent of the sample size.

## [Efficient Active Learning with Abstention](#)

- Yinglun Zhu · Robert Nowak
- abstract@[open-review](#): The goal of active learning is to achieve the same accuracy achievable by passive learning, while using much fewer labels. Exponential savings in terms of label complexity have been proved in very special cases, but fundamental lower bounds show that such improvements are impossible in general. This suggests a need to explore alternative goals for active learning. Learning with abstention is one such alternative. In this setting, the active learning algorithm may abstain from prediction and incur an error that is marginally smaller than random guessing. We develop the first computationally efficient active learning algorithm with abstention. Our algorithm provably achieves  $\mathsf{polylog}(\frac{1}{\epsilon})$  label complexity, without any low noise conditions. Such performance guarantee reduces the label complexity by an exponential factor, relative to passive learning and/or active learning that is not allowed to abstain. Furthermore, our algorithm is guaranteed to only abstain on hard examples (where the true label distribution is close to a fair coin), a novel property we term *proper abstention* that also leads to a host of other desirable characteristics (e.g., recovering minimax guarantees in the standard setting, and avoiding the undesirable ``noise-seeking'' behavior often seen in active learning). We also provide novel extensions of our algorithm that achieve *constant* label complexity and deal with model misspecification.

## [Bessel Equivariant Networks for Inversion of Transmission Effects in Multi-Mode Optical Fibres](#)

- Joshua Mitton · Simon Mkhail · Miles Padgett · Daniele Faccio · Marco Aversa · Roderick Murray-Smith
- abstract@[open-review](#): We develop a new type of model for solving the task of inverting the transmission effects of multi-mode optical fibres through the construction of an  $\mathrm{SO}^{+}(2,1)$ -equivariant neural network. This model takes advantage of the azimuthal correlations known to exist in fibre speckle patterns and naturally accounts for the difference in spatial arrangement between input and speckle patterns. In addition, we use a second post-processing network to remove circular artifacts, fill gaps, and sharpen the images, which is required due to the nature of optical fibre transmission. This two stage approach allows for the inspection of the predicted images produced by the more robust physically motivated equivariant model, which could be useful in a safety-critical application, or by the output of both models, which produces high quality images. Further, this model can scale to previously unachievable resolutions of imaging with multi-mode optical fibres and is demonstrated on  $256 \times 256$  pixel images. This is a result of improving the trainable parameter requirement from  $\mathcal{O}(N^4)$  to  $\mathcal{O}(m)$ , where  $N$  is pixel size and  $m$  is number of fibre modes. Finally, this model generalises to new images, outside of the set of training data classes, better than previous models.

## [Scalable Distributional Robustness in a Class of Non-Convex Optimization with Guarantees](#)

- Avinandan Bose · Arunesh Sinha · Tien Mai
- abstract@[open-review](#): Distributionally robust optimization (DRO) has shown a lot of promise in providing robustness in learning as well as sample-based optimization problems. We endeavor to provide DRO solutions for a class of sum of fractionals, non-convex optimization which is used for decision making in prominent areas such as facility location and security games. In contrast to previous work, we find it more tractable to optimize the equivalent variance regularized form of DRO rather than the minimax form. We transform the variance regularized form to a mixed-integer second-order cone program (MISOCP), which, while guaranteeing global optimality, does not scale enough to solve problems with real-world datasets. We further propose two abstraction approaches based on clustering and stratified sampling to increase scalability, which we then use for real-world datasets. Importantly, we provide global optimality guarantees for our approach and show experimentally that our solution quality is better than the locally optimal ones achieved by state-of-the-art gradient-based methods. We experimentally compare our different approaches and baselines and reveal nuanced properties of a DRO solution.

## [CLIPDraw: Exploring Text-to-Drawing Synthesis through Language-Image Encoders](#)

- Kevin Frans · Olaf Witkowski · Lisa Soros
- abstract@[open-review](#): CLIPDraw is an algorithm that synthesizes novel drawings from natural language input. It does not require any additional training; rather, a pre-trained CLIP language-image encoder is used as a metric for maximizing similarity between the given description and a generated drawing. Crucially, CLIPDraw operates over vector strokes rather than pixel images, which biases drawings towards simpler human-recognizable shapes. Results compare CLIPDraw with other synthesis-through-optimization methods, as well as highlight various interesting behaviors of CLIPDraw.

## [Reconstructing Training Data From Trained Neural Networks](#)

- Niv Haim · Gal Vardi · Gilad Yehudai · Michal Irani · Ohad Shamir
- abstract@[open-review](#): Understanding to what extent neural networks memorize training data is an intriguing question with practical and theoretical implications. In this paper we show that in some cases a significant fraction of the training data can in fact be reconstructed from the parameters of a trained neural network classifier. We propose a novel reconstruction scheme that stems from recent theoretical results about the implicit bias in training neural networks with gradient-based methods. To the best of our knowledge, our results are the first to show that reconstructing a large portion of the actual

training samples from a trained neural network classifier is generally possible. This has negative implications on privacy, as it can be used as an attack for revealing sensitive training data. We demonstrate our method for binary MLP classifiers on a few standard computer vision datasets.

## [If Influence Functions are the Answer, Then What is the Question?](#)

- Juhan Bae · Nathan Ng · Alston Lo · Marzyeh Ghassemi · Roger Grosse
- abstract@[open-review](#): Influence functions efficiently estimate the effect of removing a single training data point on a model's learned parameters. While influence estimates align well with leave-one-out retraining for linear models, recent works have shown this alignment is often poor in neural networks. In this work, we investigate the specific factors that cause this discrepancy by decomposing it into five separate terms. We study the contributions of each term on a variety of architectures and datasets and how they vary with factors such as network width and training time. While practical influence function estimates may be a poor match to leave-one-out retraining for nonlinear networks, we show that they are often a good approximation to a different object we term the proximal Bregman response function (PBRF). Since the PBRF can still be used to answer many of the questions motivating influence functions, such as identifying influential or mislabeled examples, our results suggest that current algorithms for influence function estimation give more informative results than previous error analyses would suggest.

## [InsNet: An Efficient, Flexible, and Performant Insertion-based Text Generation Model](#)

- Sidi Lu · Tao Meng · Nanyun Peng
- abstract@[open-review](#): We propose InsNet, an expressive insertion-based text generator with efficient training and flexible decoding (parallel or sequential). Unlike most existing insertion-based text generation works that require re-encoding of the (decoding) context after each insertion operation and thus are inefficient to train, InsNet only requires one pass of context encoding for the entire insertion sequence during training by using a novel insertion-oriented position encoding to enable computation sharing. Furthermore, InsNet provides a controllable switch between parallel and sequential decoding, making it flexible to handle more parallelizable tasks such as machine translation to support efficient decoding, or less parallelizable tasks such as lexically constrained text generation to guarantee high-quality outputs. Experiments on two unsupervised lexically constrained text generation datasets and three machine translation datasets demonstrate InsNet's advantages over previous insertion-based methods in terms of training speed, inference efficiency, and generation quality.

## [A Single Self-Supervised Model for Many Speech Modalities Enables Zero-Shot Modality Transfer](#)

- Wei-Ning Hsu · Bowen Shi
- abstract@[open-review](#): While audio-visual speech models can yield superior performance and robustness compared to audio-only models, their development and adoption are hindered by the lack of labeled and unlabeled audio-visual data and the cost to deploy one model per modality. In this paper, we present u-HuBERT, a self-supervised pre-training framework that can leverage both multimodal and unimodal speech with a unified masked cluster prediction objective. By utilizing modality dropout during pre-training, we demonstrate that a single fine-tuned model can achieve performance on par or better than the state-of-the-art modality-specific models. Moreover, our model fine-tuned only on audio can perform well with audio-visual and visual speech input, achieving zero-shot modality generalization for speech recognition and speaker verification. In particular, our single model yields 1.2%/1.4%/27.2% speech recognition word error rate on LRS3 with audio-visual/audio/visual input.

## [Expected Improvement for Contextual Bandits](#)

- Hung Tran-The · Sunil Gupta · Santu Rana · Tuan Truong · Long Tran-Thanh · Svetha Venkatesh
- abstract@[open-review](#): The expected improvement (EI) is a popular technique to handle the tradeoff between exploration and exploitation under uncertainty. However, compared to other techniques as Upper Confidence Bound (UCB) and Thompson Sampling (TS), the theoretical properties of EI have not been well studied even for non-contextual settings such as standard bandit and Bayesian optimization. In this paper, we introduce and study the EI technique as a new tool for the contextual bandit problem which is a generalization of the standard bandit. We propose two novel EI based algorithms for this problem, one when the reward function is assumed to be linear and the other for more general reward functions. With a linear reward function, we demonstrate that our algorithm achieves a near-optimal regret. Our regret improves that of LinTS \cite{agrawal13} by a factor  $\sqrt{d}$  while avoiding to solve a NP-hard problem at each iteration as in LinUCB \cite{abbasi11}. For more general reward functions, we use deep neural networks to model the reward function, and prove that our algorithm achieves a  $\tilde{O}(\tilde{d}\sqrt{T})$  regret, where  $\tilde{d}$  is the ``effective'' dimension of a neural tangent kernel matrix and  $T$  is the number of iterations. Finally, we provide an empirical evaluation of the algorithms on various benchmark datasets. Our experiments show that our algorithms work well and consistently outperform existing approaches, especially in high dimensions.

## [Toward Understanding Privileged Features Distillation in Learning-to-Rank](#)

- Shuo Yang · Sujay Sanghavi · Holakou Rahmani · Jan Bakus · Vishwanathan S. V. N.
- abstract@[open-review](#): In learning-to-rank problems, a \textit{privileged feature} is one that is available during model training, but not available at test time. Such features naturally arise in merchandised recommendation systems; for instance, "user clicked this item" as a feature is predictive of "user purchased this item" in the offline data, but is clearly not available during online serving. Another source of privileged features is those that are too expensive to compute online but feasible to be added offline. \textit{Privileged features distillation} (PFD) refers to a natural idea: train a "teacher" model using all features (including privileged ones) and then use it to train a "student" model that does not use the privileged features. In this paper, we first study PFD empirically on three public ranking datasets and an industrial-scale ranking problem derived from Amazon's logs. We show that PFD outperforms several baselines (no-distillation, pretraining-finetuning, self-distillation, and generalized distillation) on all these datasets. Next, we analyze why and when PFD performs well via both empirical ablation studies and theoretical analysis for linear models. Both investigations uncover an interesting non-monotone behavior: as the predictive power of a privileged feature increases, the performance of the resulting student model initially increases but then decreases. We show the reason for the later decreasing performance is that a very predictive privileged teacher produces predictions with high variance, which lead to high variance student estimates and inferior testing performance.

## [Hyperbolic Feature Augmentation via Distribution Estimation and Infinite Sampling on Manifolds](#)

- Zhi Gao · Yuwei Wu · Yunde Jia · Mehrtash Harandi
- abstract@[open-review](#): Learning in the hyperbolic space has attracted growing attention recently, owing to its high capability in capturing hierarchical structures. However, existing learning algorithms in the hyperbolic space tend to overfit when limited data is given. In this paper, we propose a hyperbolic feature augmentation method that generates diverse and discriminative features in the hyperbolic space to combat overfitting. We employ wrapped hyperbolic normal distributions to model augmented features, and use a neural ordinary differential equation (ODE) module that benefits from meta-learning to estimate the distribution. In this case, the bias of estimation caused by the scarcity of data is reduced. We also derive an upper bound of the augmentation loss, which enables us to train a hyperbolic model by using an infinite number of augmentations. Experiments on few-shot learning and continual learning tasks show that our method significantly improves the performance of hyperbolic algorithms in low data regimes.

## [Pushing the limits of fairness impossibility: Who's the fairest of them all?](#)

- Brian Hsu · Rahul Mazumder · Preetam Nandy · Kinjal Basu

- abstract@[open-review](#): The impossibility theorem of fairness is a foundational result in the algorithmic fairness literature. It states that outside of special cases, one cannot exactly and simultaneously satisfy all three common and intuitive definitions of fairness - demographic parity, equalized odds, and predictive rate parity. This result has driven most works to focus on solutions for one or two of the metrics. Rather than follow suit, in this paper we present a framework that pushes the limits of the impossibility theorem in order to satisfy all three metrics to the best extent possible. We develop an integer-programming based approach that can yield a certifiably optimal post-processing method for simultaneously satisfying multiple fairness criteria under small violations. We show experiments demonstrating that our post-processor can improve fairness across the different definitions simultaneously with minimal model performance reduction. We also discuss applications of our framework for model selection and fairness explainability, thereby attempting to answer the question: Who's the fairest of them all?

## [Spartan: Differentiable Sparsity via Regularized Transportation](#)

- Kai Sheng Tai · Taipeng Tian · Ser Nam Lim
- abstract@[open-review](#): We present Spartan, a method for training sparse neural network models with a predetermined level of sparsity. Spartan is based on a combination of two techniques: (1) soft top-k masking of low-magnitude parameters via a regularized optimal transportation problem and (2) dual averaging-based parameter updates with hard sparsification in the forward pass. This scheme realizes an exploration-exploitation tradeoff: early in training, the learner is able to explore various sparsity patterns, and as the soft top-k approximation is gradually sharpened over the course of training, the balance shifts towards parameter optimization with respect to a fixed sparsity mask. Spartan is sufficiently flexible to accommodate a variety of sparsity allocation policies, including both unstructured and block-structured sparsity, global and per-layer sparsity budgets, as well as general cost-sensitive sparsity allocation mediated by linear models of per-parameter costs. On ImageNet-1K classification, we demonstrate that training with Spartan yields 95% sparse ResNet-50 models and 90% block sparse ViT-B/16 models while incurring absolute top-1 accuracy losses of less than 1% compared to fully dense training.

## [Posted Pricing and Dynamic Prior-independent Mechanisms with Value Maximizers](#)

- Yuan Deng · Vahab Mirrokni · Hanrui Zhang
- abstract@[open-review](#): We study posted price auctions and dynamic prior-independent mechanisms for (ROI-constrained) value maximizers. In contrast to classic (quasi-linear) utility maximizers, these agents aim to maximize their total value subject to a minimum ratio of value per unit of payment made. When personalized posted prices are allowed, posted price auctions for value maximizers can be reduced to posted price auctions for utility maximizers. However, for anonymous posted prices, the well-known  $\frac{1}{2}$  approximation for utility maximizers is impossible for value maximizers and we provide a posted price mechanism with  $\frac{1}{12}(1 - 1/e)$  approximation. Moreover, we demonstrate how to apply our results to design prior-independent mechanisms in a dynamic environment; and to the best of our knowledge, this gives the first constant revenue approximation with multiple value maximizers. Finally, we provide an extension to combinatorial auctions with submodular / XOS agents.

## [Provably tuning the ElasticNet across instances](#)

- Maria-Florina Balcan · Misha Khodak · Dravyansh Sharma · Ameet Talwalkar
- abstract@[open-review](#): An important unresolved challenge in the theory of regularization is to set the regularization coefficients of popular techniques like the ElasticNet with general provable guarantees. We consider the problem of tuning the regularization parameters of Ridge regression, LASSO, and the ElasticNet across multiple problem instances, a setting that encompasses both cross-validation and multi-task hyperparameter optimization. We obtain a novel structural result for the ElasticNet which characterizes the loss as a piecewise rational function of the tuning parameters, with algebraic boundaries. We use this to bound the structural complexity of the regularized loss functions, and show generalization guarantees for tuning the ElasticNet regression coefficients in the statistical setting. We also consider the more challenging online learning setting, and show vanishing average expected regret relative to the optimal parameter pair. We also extend our results to tuning classification algorithms obtained by thresholding regression fits regularized by Ridge, LASSO or ElasticNet. Our results are the first general learning-theoretic guarantees, without strong assumptions on the data distribution, for this important class of problems. Our guarantees hold for both validation and popular information criterion objectives.

## [Causality Preserving Chaotic Transformation and Classification using Neurochaos Learning](#)

- Harikrishnan N B · Aditi Kathpalia · Nithin Nagaraj
- abstract@[open-review](#): Discovering cause-effect from observational data is an important but challenging problem in science and engineering. In this work, a recently proposed brain inspired learning algorithm namely-\textit{Neurochaos Learning} (NL) is used for the classification of cause-effect from coupled autoregressive processes, coupled 1D chaotic skew tent maps, coupled 1D chaotic logistic maps and a real-world prey-predator system. In the case of coupled skew tent maps, the proposed method consistently outperforms a five layer Deep Neural Network (DNN) and Long Short Term Memory (LSTM) architecture for coupling coefficient values ranging from \$0.1\$ to \$0.7\$. Further, we investigate the preservation of causality in the feature extracted space of NL using Granger Causality for coupled autoregressive processes and Compression-Complexity Causality for coupled chaotic systems and real-world prey-predator dataset. Unlike DNN, LSTM and 1D Convolutional Neural Network, it is found that NL preserves the inherent causal structures present in the input timeseries data. This finding is promising for the theory and applications of causal machine learning and opens up the possibility to explore the potential of NL for more sophisticated causal learning tasks.

## [A First Approach to Universal Second-Order Acceleration for Convex Minimization](#)

- Ali Kavis · Kimon Antonakopoulos · Volkan Cevher
- abstract@[open-review](#): In this work, we propose a universal and adaptive second-order method for minimization of second-order smooth, convex functions. Precisely, our algorithm achieves  $O(\sigma / \sqrt{T})$  when the oracle feedback is stochastic with variance  $\sigma$ , and obtains the improved  $O(1 / T^3)$  convergence with deterministic oracles. Our method achieves this rate interpolation without knowing the nature of the oracle a priori, which was enabled by a parameter-free step-size that is oblivious to the knowledge of smoothness modulus, variance bounds and the diameter of the constrained set. To our knowledge, this is the first universal algorithm that achieves the aforementioned global guarantees within second-order convex optimization literature.

## [Trimmed Maximum Likelihood Estimation for Robust Generalized Linear Model](#)

- Weihao Kong · Rajat Sen · Pranjal Awasthi · Abhimanyu Das
- abstract@[open-review](#): We study the problem of learning generalized linear models under adversarial corruptions. We analyze a classical heuristic called the \textit{iterative trimmed maximum likelihood estimator} which is known to be effective against \textit{label corruptions} in practice. Under label corruptions, we prove that this simple estimator achieves minimax near-optimal risk on a wide range of generalized linear models, including Gaussian regression, Poisson regression and Binomial regression. Finally, we extend the estimator to the much more challenging setting of \textit{label and covariate corruptions} and demonstrate its robustness and optimality in that setting as well.

## [Outsourcing Training without Uploading Data via Efficient Collaborative Open-Source Sampling](#)

- Junyuan Hong · Lingjuan Lyu · Jiayu Zhou · Michael Spranger

- abstract@[open-review](#): As deep learning blooms with growing demand for computation and data resources, outsourcing model training to a powerful cloud server becomes an attractive alternative to training at a low-power and cost-effective end device. Traditional outsourcing requires uploading device data to the cloud server, which can be infeasible in many real-world applications due to the often sensitive nature of the collected data and the limited communication bandwidth. To tackle these challenges, we propose to leverage widely available open-source data, which is a massive dataset collected from public and heterogeneous sources (e.g., Internet images). We develop a novel strategy called Efficient Collaborative Open-source Sampling (ECOS) to construct a proximal proxy dataset from open-source data for cloud training, in lieu of client data. ECOS probes open-source data on the cloud server to sense the distribution of client data via a communication- and computation-efficient sampling process, which only communicates a few compressed public features and client scalar responses. Extensive empirical studies show that the proposed ECOS improves the quality of automated client labeling, model compression, and label outsourcing when applied in various learning scenarios. Source codes will be released.

## [FNeVR: Neural Volume Rendering for Face Animation](#)

- Bohan Zeng · Boyu Liu · Hong Li · Xuhui Liu · Jianzhuang Liu · Dapeng Chen · Wei Peng · Baochang Zhang
- abstract@[open-review](#): Face animation, one of the hottest topics in computer vision, has achieved a promising performance with the help of generative models. However, it remains a critical challenge to generate identity preserving and photo-realistic images due to the sophisticated motion deformation and complex facial detail modeling. To address these problems, we propose a Face Neural Volume Rendering (FNeVR) network to fully explore the potential of 2D motion warping and 3D volume rendering in a unified framework. In FNeVR, we design a 3D Face Volume Rendering (FVR) module to enhance the facial details for image rendering. Specifically, we first extract 3D information with a well designed architecture, and then introduce an orthogonal adaptive ray-sampling module for efficient rendering. We also design a lightweight pose editor, enabling FNeVR to edit the facial pose in a simple yet effective way. Extensive experiments show that our FNeVR obtains the best overall quality and performance on widely used talking-head benchmarks.

## [Rethinking Variational Inference for Probabilistic Programs with Stochastic Support](#)

- Tim Reichelt · Luke Ong · Thomas Rainforth
- abstract@[open-review](#): We introduce Support Decomposition Variational Inference (SDVI), a new variational inference (VI) approach for probabilistic programs with stochastic support. Existing approaches to this problem rely on designing a single global variational guide on a variable-by-variable basis, while maintaining the stochastic control flow of the original program. SDVI instead breaks the program down into sub-programs with static support, before automatically building separate sub-guides for each. This decomposition significantly aids in the construction of suitable variational families, enabling, in turn, substantial improvements in inference performance.

## [Low-rank Optimal Transport: Approximation, Statistics and Debiasing](#)

- Meyer Scetbon · Marco Cuturi
- abstract@[open-review](#): The matching principles behind optimal transport (OT) play an increasingly important role in machine learning, a trend which can be observed when OT is used to disambiguate datasets in applications (e.g. single-cell genomics) or used to improve more complex methods (e.g. balanced attention in transformers or self-supervised learning). To scale to more challenging problems, there is a growing consensus that OT requires solvers that can operate on millions, not thousands, of points. The low-rank optimal transport (LOT) approach advocated in \cite{scetbon2021lowrank} holds several promises in that regard, and was shown to complement more established entropic regularization approaches, being able to insert itself in more complex pipelines, such as quadratic OT. LOT restricts the search for low-cost couplings to those that have a low-nonnegative rank, yielding linear time algorithms in cases of interest. However, these promises can only be fulfilled if the LOT approach is seen as a legitimate contender to entropic regularization when compared on properties of interest, where the scorecard typically includes theoretical properties (statistical bounds, relation to other methods) or practical aspects (debiasing, hyperparameter tuning, initialization). We target each of these areas in this paper in order to cement the impact of low-rank approaches in computational OT.

## [Finite-Time Last-Iterate Convergence for Learning in Multi-Player Games](#)

- Yang Cai · Argyris Oikonomou · Weiqiang Zheng
- abstract@[open-review](#): We study the question of last-iterate convergence rate of the extragradient algorithm by Korpelevich [1976] and the optimistic gradient algorithm by Popov [1980] in multi-player games. We show that both algorithms with constant step-size have last-iterate convergence rate of  $\$O(\frac{1}{\sqrt{T}})$  to a Nash equilibrium in terms of the gap function in smooth monotone games, where each player's action set is an arbitrary convex set. Previous results only study the unconstrained setting, where each player's action set is the entire Euclidean space. Our results address an open question raised in several recent work by Hsieh et al. [2019], Golowich et al. [2020a,b], who ask for last-iterate convergence rate of either the extragradient or the optimistic gradient algorithm in the constrained setting. Our convergence rates for both algorithms are tight and match the lower bounds by Golowich et al. [2020a,b]. At the core of our results lies a new notion -- the tangent residual, which we use to measure the proximity to equilibrium. We use the tangent residual (or a slight variation of the tangent residual) as the the potential function in our analysis of the extragradient algorithm (or the optimistic gradient algorithm) and prove that it is non-increasing between two consecutive iterates.

## [Non-rigid Point Cloud Registration with Neural Deformation Pyramid](#)

- YANG LI · Tatsuya Harada
- abstract@[open-review](#): Non-rigid point cloud registration is a key component in many computer vision and computer graphics applications. The high complexity of the unknown non-rigid motion make this task a challenging problem. In this paper, we break down this problem via hierarchical motion decomposition. Our method called Neural Deformation Pyramid (NDP) represents non-rigid motion using a pyramid architecture. Each pyramid level, denoted by a Multi-Layer Perception (MLP), takes as input a sinusoidally encoded 3D point and outputs its motion increments from the previous level. The sinusoidal function starts with a low input frequency and gradually increases when the pyramid level goes down. This allows a multi-level rigid to nonrigid motion decomposition and also speeds up the solving by ~50 times compared to the existing MLP-based approach. Our method achieves advanced partial-to-partial non-rigid point cloud registration results on the 4DMatch/4DLoMatchbenchmark under both no-learned and supervised settings.

## [Efficient Methods for Non-stationary Online Learning](#)

- Peng Zhao · Yan-Feng Xie · Lijun Zhang · Zhi-Hua Zhou
- abstract@[open-review](#): Non-stationary online learning has drawn much attention in recent years. In particular, \emph{dynamic regret} and \emph{adaptive regret} are proposed as two principled performance measures for online convex optimization in non-stationary environments. To optimize them, a two-layer online ensemble is usually deployed due to the inherent uncertainty of the non-stationarity, in which a group of base-learners are maintained and a meta-algorithm is employed to track the best one on the fly. However, the two-layer structure raises the concern about the computational complexity --- those methods typically maintain  $O(\log T)$  base-learners simultaneously for a  $T$ -round online game and thus perform multiple projections onto the feasible domain per round, which becomes the computational bottleneck when the domain is complicated. In this paper, we present efficient methods for optimizing dynamic regret and adaptive regret, which reduce the number of projections per round from  $O(\log T)$  to 1. Moreover, our obtained algorithms require only one gradient query and one function evaluation at each round. Our technique hinges on the reduction mechanism developed in parameter-free online learning and requires non-trivial twists on non-stationary online methods. Empirical studies verify our theoretical findings.

## Constrained Langevin Algorithms with L-mixing External Random Variables

- Yuping Zheng · Andrew Lamperski
- abstract@[open-review](#): Langevin algorithms are gradient descent methods augmented with additive noise, and are widely used in Markov Chain Monte Carlo (MCMC) sampling, optimization, and learning. In recent years, the non-asymptotic analysis of Langevin algorithms for non-convex optimization learning has been extensively explored. For constrained problems with non-convex losses over compact convex domain in the case of IID data variables, Langevin algorithm achieves a deviation of  $\mathcal{O}(T^{-1/4} (\log T)^{1/2})$  from its target distribution [\[lamperski2021projected\]](#). In this paper, we obtain a deviation of  $\mathcal{O}(T^{-1/2} \log T)$  in  $\mathbb{W}$ -Wasserstein distance for non-convex losses with  $L$ -mixing data variables and polyhedral constraints (which are not necessarily bounded). This deviation indicates that our convergence rate is faster than those in the previous works on constrained Langevin algorithms for non-convex optimization.

## ASPiRe: Adaptive Skill Priors for Reinforcement Learning

- Mengda Xu · Manuela Veloso · Shuran Song
- abstract@[open-review](#): We introduce ASpiRe (Adaptive Skill Prior for RL), a new approach that leverages prior experience to accelerate reinforcement learning. Unlike existing methods that learn a single skill prior from a large and diverse dataset, our framework learns a library of different distinction skill priors (i.e., behavior priors) from a collection of specialized datasets, and learns how to combine them to solve a new task. This formulation allows the algorithm to acquire a set of specialized skill priors that are more reusable for downstream tasks; however, it also brings up additional challenges of how to effectively combine these unstructured sets of skill priors to form a new prior for new tasks. Specifically, it requires the agent not only to identify which skill prior(s) to use but also how to combine them (either sequentially or concurrently) to form a new prior. To achieve this goal, ASpiRe includes Adaptive Weight Module (AWM) that learns to infer an adaptive weight assignment between different skill priors and uses them to guide policy learning for downstream tasks via weighted Kullback-Leibler divergences. Our experiments demonstrate that ASpiRe can significantly accelerate the learning of new downstream tasks in the presence of multiple priors and show improvement on competitive baselines.

## Policy Gradient With Serial Markov Chain Reasoning

- Edoardo Cetin · Oya Celiktutan
- abstract@[open-review](#): We introduce a new framework that performs decision-making in reinforcement learning (RL) as an iterative reasoning process. We model agent behavior as the steady-state distribution of a parameterized reasoning Markov chain (RMC), optimized with a new tractable estimate of the policy gradient. We perform action selection by simulating the RMC for enough reasoning steps to approach its steady-state distribution. We show our framework has several useful properties that are inherently missing from traditional RL. For instance, it allows agent behavior to approximate any continuous distribution over actions by parameterizing the RMC with a simple Gaussian transition function. Moreover, the number of reasoning steps to reach convergence can scale adaptively with the difficulty of each action selection decision and can be accelerated by re-using past solutions. Our resulting algorithm achieves state-of-the-art performance in popular Mujoco and DeepMind Control benchmarks, both for proprioceptive and pixel-based tasks.

## Fine-Grained Analysis of Stability and Generalization for Modern Meta Learning Algorithms

- Jiechao Guan · Yong Liu · Zhiwu Lu
- abstract@[open-review](#): The support/query episodic training strategy has been widely applied in modern meta learning algorithms. Supposing the  $n$  training episodes and the test episodes are sampled independently from the same environment, previous work has derived a generalization bound of  $\mathcal{O}(1/\sqrt{n})$  for smooth non-convex functions via algorithmic stability analysis. In this paper, we provide fine-grained analysis of stability and generalization for modern meta learning algorithms by considering more general situations. Firstly, we develop matching lower and upper stability bounds for meta learning algorithms with two types of loss functions: (1) nonsmooth convex functions with  $\alpha$ -Hölder continuous subgradients ( $\alpha \in [0,1]$ ); (2) smooth (including convex and non-convex) functions. Our tight stability bounds show that, in the nonsmooth convex case, meta learning algorithms can be inherently less stable than in the smooth convex case. For the smooth non-convex functions, our stability bound is sharper than the existing one, especially in the setting where the number of iterations is larger than the number  $n$  of training episodes. Secondly, we derive improved generalization bounds for meta learning algorithms that hold with high probability. Specifically, we first demonstrate that, under the independent episode environment assumption, the generalization bound of  $\mathcal{O}(1/\sqrt{n})$  via algorithmic stability analysis is near optimal. To attain faster convergence rate, we show how to yield a deformed generalization bound of  $\mathcal{O}(\ln n/n)$  with the curvature condition of loss functions. Finally, we obtain a generalization bound for meta learning with dependent episodes whose dependency relation is characterized by a graph. Experiments on regression problems are conducted to verify our theoretical results.

## Semantic Difference Convolution for Semantic Segmentation

- Haoru Tan · Sitong Wu · Jimin Pi
- abstract@[open-review](#): Precise and accurate segmentation over boundary areas is important in semantic segmentation. The commonly used convolutional operators tend to smooth and blur local detail cues, making it difficult for deep learning models to generate accurate boundary predictions. In this paper, we propose an efficient boundary-aware convolution operator to boost the boundary modeling capacity for semantic segmentation, named Semantic Difference Convolution (SDC). The SDC is sensitive to the inter-class boundary, while ignoring the noisy intra-class pseudo-boundaries. Based on the SDC operator, we further design a lightweight module, termed Semantic Difference Module (SDM) to enhance the boundary-related information. The SDM can be flexibly plugged into any existing encoder-decoder segmentation model. Extensive experiments show that our approach can achieve consistent improvements (especially for boundary regions) over several typical state-of-the-art segmentation baseline models on four challenging benchmarks, including ADE20K, Cityscapes, COCO-Stuff, and PASCAL-Context.

## Is a Modular Architecture Enough?

- Sarthak Mittal · Yoshua Bengio · Guillaume Lajoie
- abstract@[open-review](#): Inspired from human cognition, machine learning systems are gradually revealing advantages of sparser and more modular architectures. Recent work demonstrates that not only do some modular architectures generalize well, but they also lead to better out of distribution generalization, scaling properties, learning speed, and interpretability. A key intuition behind the success of such systems is that the data generating system for most real-world settings is considered to consist of sparse modular connections, and endowing models with similar inductive biases will be helpful. However, the field has been lacking in a rigorous quantitative assessment of such systems because these real-world data distributions are complex and unknown. In this work, we provide a thorough assessment of common modular architectures, through the lens of simple and known modular data distributions. We highlight the benefits of modularity and sparsity and reveal insights on the challenges faced while optimizing modular systems. In doing so, we propose evaluation metrics that highlight the benefits of modularity, the regimes in which these benefits are substantial, as well as the sub-optimality of current end-to-end learned modular systems as opposed to their claimed potential.

## Queue Up Your Regrets: Achieving the Dynamic Capacity Region of Multiplayer Bandits

- Ilai Bistritz · Nicholas Bambos

- abstract@[open-review](#): Consider  $\$N$  cooperative agents such that for  $T$  turns, each agent  $n$  takes an action  $a_n$  and receives a stochastic reward  $r_n(a_1, \dots, a_N)$ . Agents cannot observe the actions of other agents and do not know even their own reward function. The agents can communicate with their neighbors on a connected graph  $G$  with diameter  $d(G)$ . We want each agent  $n$  to achieve an expected average reward of at least  $\lambda_n$  over time, for a given quality of service (QoS) vector  $\boldsymbol{\lambda}$ . A QoS vector  $\boldsymbol{\lambda}$  is not necessarily achievable. By giving up on immediate reward, knowing that the other agents will compensate later, agents can improve their achievable capacity region. Our main observation is that the gap between  $\lambda_n$  and the accumulated reward of agent  $n$ , which we call the QoS regret, behaves like a queue. Inspired by this observation, we propose a distributed algorithm that aims to learn a max-weight matching of agents to actions. In each epoch, the algorithm employs a consensus phase where the agents agree on a certain weighted sum of rewards by communicating only  $O(d(G))$  numbers every turn. Then, the algorithm uses distributed successive elimination on a random subset of action profiles to approximately maximize this weighted sum of rewards. We prove a bound on the sum of expected QoS regrets of all agents, that holds if  $\boldsymbol{\lambda}$  is a safety margin  $\epsilon_T$  away from the boundary of the capacity region, where  $\epsilon_T \rightarrow 0$  as  $T \rightarrow \infty$ . This bound implies that, for large  $T$ , our algorithm can achieve any  $\boldsymbol{\lambda}$  in the interior of the dynamic capacity region, with an empirical average expected QoS regret of  $\tilde{O}(\sqrt{t})$  over  $t=1, \dots, T$  which never exceeds  $\tilde{O}(\sqrt{t})$  for any  $t$ . We then extend our result to the case of a time-varying i.i.d. communication graph.

## [Trading Off Resource Budgets For Improved Regret Bounds](#)

- Thomas Orton · Damon Falck
- abstract@[open-review](#): We consider a variant of adversarial online learning where in each round one picks  $B$  arms and incurs cost equal to the minimum of the costs of each arm chosen. We study the trade-off between the budget  $B$  and the regret  $R_T^*$  relative to the best fixed arm in hindsight. By adapting the techniques of Kalai and Vempala [2005], we show that in the full feedback setting Following the Top  $B$  perturbed Leaders (FTL) achieves expected regret  $\mathcal{O}(T^{1/(B+1)} \ln(N)^{1/(B+1)})$  with  $N$  arms over time horizon  $T$ , and allows one to trade off an arm budget for improved regret bounds. We observe that algorithms which use standard regret minimizers as subroutines can sometimes be adapted by replacing these subroutines with FTL. We use this observation to generalize existing algorithms for Online Submodular Function Maximization [Streeter and Golovin, 2008] in both the full feedback and semi-bandit feedback settings to allow trade-offs between resource budgets and regret bounds, and we empirically evaluate these new algorithms on an online hyperparameter optimization problem. Likewise, we show how FTL can lead to new algorithms for Linear Programming which require stronger oracles at the benefit of fewer oracle calls.

## [Neural Matching Fields: Implicit Representation of Matching Cost for Semantic Correspondence](#)

- Sungewan Hong · Seungryong Kim · Dongbo Min · Sangryul Jeon · Seokju Cho · Susung Hong · Jisu Nam
- abstract@[open-review](#): Existing pipelines of semantic correspondence commonly include extracting high-level semantic features for the invariance against intra-class variations and background clutters. This architecture, however, inevitably results in a low-resolution matching field that additionally requires an ad-hoc interpolation process as a post-processing for converting it into a high-resolution one, certainly limiting the overall performance of matching results. To overcome this, inspired by recent success of implicit neural representation, we present a novel method for semantic correspondence, called neural matching field (NeMF). However, complicity and high-dimensionality of a 4D matching field are the major hindrances. To address them, we propose a cost embedding network consisting of convolution and self-attention layers to process the coarse cost volume to obtain cost feature representation, which is used as a guidance for establishing high-precision matching field through the following fully-connected network. Although this may help to better structure the matching field, learning a high-dimensional matching field remains challenging mainly due to computational complexity, since a naive exhaustive inference would require querying from all pixels in the 4D space to infer pixel-wise correspondences. To overcome this, in the training phase, we randomly sample matching candidates. In the inference phase, we propose a novel inference approach which iteratively performs PatchMatch-based inference and coordinate optimization at test time. With the proposed method, state-of-the-art performance is attained on several standard benchmarks for semantic correspondence.

## [Batch-Size Independent Regret Bounds for Combinatorial Semi-Bandits with Probabilistically Triggered Arms or Independent Arms](#)

- Xutong Liu · Jinhang Zuo · Siwei Wang · Carlee Joe-Wong · John C.S. Lui · Wei Chen
- abstract@[open-review](#): In this paper, we study the combinatorial semi-bandits (CMAB) and focus on reducing the dependency of the batch-size  $K$  in the regret bound, where  $K$  is the total number of arms that can be pulled or triggered in each round. First, for the setting of CMAB with probabilistically triggered arms (CMAB-T), we discover a novel (directional) triggering probability and variance modulated (TPVM) condition that can replace the previously-used smoothness condition for various applications, such as cascading bandits, online network exploration and online influence maximization. Under this new condition, we propose a BCUCB-T algorithm with variance-aware confidence intervals and conduct regret analysis which reduces the  $\mathcal{O}(K)$  factor to  $\mathcal{O}(\log K)$  or  $\mathcal{O}(\log^2 K)$  in the regret bound, significantly improving the regret bounds for the above applications. Second, for the setting of non-triggering CMAB with independent arms, we propose a SESCB algorithm which leverages on the non-triggering version of the TPVM condition and completely removes the dependency on  $K$  in the leading regret. As a valuable by-product, the regret analysis used in this paper can improve several existing results by a factor of  $\mathcal{O}(\log K)$ . Finally, experimental evaluations show our superior performance compared with benchmark algorithms in different applications.

## [An Embarrassingly Simple Approach to Semi-Supervised Few-Shot Learning](#)

- Xiu-Shen Wei · H.-Y. Xu · Faen Zhang · Yuxin Peng · Wei Zhou
- abstract@[open-review](#): Semi-supervised few-shot learning consists in training a classifier to adapt to new tasks with limited labeled data and a fixed quantity of unlabeled data. Many sophisticated methods have been developed to address the challenges this problem comprises. In this paper, we propose a simple but quite effective approach to predict accurate negative pseudo-labels of unlabeled data from an indirect learning perspective, and then augment the extremely label-constrained support set in few-shot classification tasks. Our approach can be implemented in just few lines of code by only using off-the-shelf operations, yet it is able to outperform state-of-the-art methods on four benchmark datasets.

## [Rethinking and Scaling Up Graph Contrastive Learning: An Extremely Efficient Approach with Group Discrimination](#)

- YIZHEN ZHENG · Shirui Pan · Vincent CS Lee · Yu Zheng · Philip S Yu
- abstract@[open-review](#): Graph contrastive learning (GCL) alleviates the heavy reliance on label information for graph representation learning (GRL) via self-supervised learning schemes. The core idea is to learn by maximising mutual information for similar instances, which requires similarity computation between two node instances. However, GCL is inefficient in both time and memory consumption. In addition, GCL normally requires a large number of training epochs to be well-trained on large-scale datasets. Inspired by an observation of a technical defect (i.e., inappropriate usage of Sigmoid function) commonly used in two representative GCL works, DGI and MVGRL, we revisit GCL and introduce a new learning paradigm for self-supervised graph representation learning, namely, Group Discrimination (GD), and propose a novel GD-based method called Graph Group Discrimination (GGD). Instead of similarity computation, GGD directly discriminates two groups of node samples with a very simple binary cross-entropy loss. In addition, GGD requires much fewer training epochs to obtain competitive performance compared with GCL methods on large-scale datasets. These two advantages endow GGD with very efficient property. Extensive experiments show that GGD outperforms state-of-the-art self-supervised methods on eight datasets. In particular, GGD can be trained in 0.18 seconds (6.44 seconds including data preprocessing) on ogbn-arxiv, which is orders of magnitude (10,000+) faster

than GCL baselines while consuming much less memory. Trained with 9 hours on ogbn-papers100M with billion edges, GGD outperforms its GCL counterparts in both accuracy and efficiency.

## [Provably sample-efficient RL with side information about latent dynamics](#)

- Yao Liu · Dipendra Misra · Miro Dudik · Robert Schapire
- abstract@[open-review](#): We study reinforcement learning (RL) in settings where observations are high-dimensional, but where an RL agent has access to abstract knowledge about the structure of the state space, as is the case, for example, when a robot is tasked to go to a specific room in a building using observations from its own camera, while having access to the floor plan. We formalize this setting as transfer reinforcement learning from an "abstract simulator," which we assume is deterministic (such as a simple model of moving around the floor plan), but which is only required to capture the target domain's latent-state dynamics approximately up to unknown (bounded) perturbations (to account for environment stochasticity). Crucially, we assume no prior knowledge about the structure of observations in the target domain except that they can be used to identify the latent states (but the decoding map is unknown). Under these assumptions, we present an algorithm, called TASID, that learns a robust policy in the target domain, with sample complexity that is polynomial in the horizon, and independent of the number of states, which is not possible without access to some prior knowledge. In synthetic experiments, we verify various properties of our algorithm and show that it empirically outperforms transfer RL algorithms that require access to "full simulators" (i.e., those that also simulate observations).

## [List-Decodable Sparse Mean Estimation](#)

- Shiwei Zeng · Jie Shen
- abstract@[open-review](#): Robust mean estimation is one of the most important problems in statistics: given a set of samples  $\{x_1, \dots, x_n\} \subset \mathbb{R}^d$  where an  $\alpha$  fraction are drawn from some distribution  $D$  and the rest are adversarially corrupted, it aims to estimate the mean of  $D$ . A surge of recent research interest has been focusing on the list-decodable setting where  $\alpha \in (0, \frac{1}{2}]$ , and the goal is to output a finite number of estimates among which at least one approximates the target mean. In this paper, we consider that the underlying distribution is Gaussian and the target mean is  $k$ -sparse. Our main contribution is the first polynomial-time algorithm that enjoys sample complexity  $O(\mathrm{poly}(k, \log d))$ , i.e. poly-logarithmic in the dimension. One of the main algorithmic ingredients is using low-degree sparse polynomials to filter outliers, which may be of independent interest.

## [Is \$L^2\$ Physics Informed Loss Always Suitable for Training Physics Informed Neural Network?](#)

- Chuwei Wang · Shanda Li · Di He · Liwei Wang
- abstract@[open-review](#): The Physics-Informed Neural Network (PINN) approach is a new and promising way to solve partial differential equations using deep learning. The  $L^2$  Physics-Informed Loss is the de-facto standard in training Physics-Informed Neural Networks. In this paper, we challenge this common practice by investigating the relationship between the loss function and the approximation quality of the learned solution. In particular, we leverage the concept of stability in the literature of partial differential equation to study the asymptotic behavior of the learned solution as the loss approaches zero. With this concept, we study an important class of high-dimensional non-linear PDEs in optimal control, the Hamilton-Jacobi-Bellman (HJB) Equation, and prove that for general  $L^p$  Physics-Informed Loss, a wide class of HJB equation is stable only if  $p$  is sufficiently large. Therefore, the commonly used  $L^2$  loss is not suitable for training PINN on those equations, while  $L^\infty$  loss is a better choice. Based on the theoretical insight, we develop a novel PINN training algorithm to minimize the  $L^\infty$  loss for HJB equations which is in a similar spirit to adversarial training. The effectiveness of the proposed algorithm is empirically demonstrated through experiments.

## [Understanding Self-Supervised Graph Representation Learning from a Data-Centric Perspective](#)

- Puja Trivedi · Ekdeep S Lubana · Mark Heimann · Danai Koutra · Jayaraman Thiagarajan
- abstract@[open-review](#): Recent analyses of self-supervised representation learning (SSL) find the following data-centric properties to be critical for learning high-quality representations: invariance to task-irrelevant semantics, separability of classes in some latent space, and recoverability of labels from augmented samples. However, given their discrete, non-Euclidean nature, graph datasets and graph SSL methods are unlikely to satisfy these properties. This raises the question: how do graph SSL methods, and in particular, contrastive learning (CL), work well? To systematically probe this question, we perform a generalization analysis for CL when using generic graph augmentations (GGAs) based on dataset recoverability and separability constraints, yielding insights into task-relevant augmentations. As we empirically show, popularly used GGAs do not induce task-relevant invariances on common benchmark datasets, leading to only marginal gains over naive, untrained baselines. Our theory motivates a synthetic data generation process that enables control over both augmentation recoverability and dataset separability, enabling a better benchmark for evaluation of graph SSL methods and identifies limitations in advanced augmentation methods. Overall, our work rigorously contextualizes, both empirically and theoretically, the effects of data-centric properties on augmentation strategies and learning paradigms for graph SSL.

## [Defending Against Adversarial Attacks via Neural Dynamic System](#)

- Xiyuan Li · Zou Xin · Weiwei Liu
- abstract@[open-review](#): Although deep neural networks (DNN) have achieved great success, their applications in safety-critical areas are hindered due to their vulnerability to adversarial attacks. Some recent works have accordingly proposed to enhance the robustness of DNN from a dynamic system perspective. Following this line of inquiry, and inspired by the asymptotic stability of the general nonautonomous dynamical system, we propose to make each clean instance be the asymptotically stable equilibrium point of a slowly time-varying system in order to defend against adversarial attacks. We present a theoretical guarantee that if a clean instance is an asymptotically stable equilibrium point and the adversarial instance is in the neighborhood of this point, the asymptotic stability will reduce the adversarial noise to bring the adversarial instance close to the clean instance. Motivated by our theoretical results, we go on to propose a nonautonomous neural ordinary differential equation (ASODE) and place constraints on its corresponding linear time-variant system to make all clean instances act as its asymptotically stable equilibrium points. Our analysis suggests that the constraints can be converted to regularizers in implementation. The experimental results show that ASODE improves robustness against adversarial attacks and outperforms the state-of-the-art methods.

## [Do We Really Need a Learnable Classifier at the End of Deep Neural Network?](#)

- Yibo Yang · Shixiang Chen · Xiangtai Li · Liang Xie · Zhouchen Lin · Dacheng Tao
- abstract@[open-review](#): Modern deep neural networks for classification usually jointly learn a backbone for representation and a linear classifier to output the logit of each class. A recent study has shown a phenomenon called neural collapse that the within-class means of features and the classifier vectors converge to the vertices of a simplex equiangular tight frame (ETF) at the terminal phase of training on a balanced dataset. Since the ETF geometric structure maximally separates the pair-wise angles of all classes in the classifier, it is natural to raise the question, why do we spend an effort to learn a classifier when we know its optimal geometric structure? In this paper, we study the potential of learning a neural network for classification with the classifier randomly initialized as an ETF and fixed during training. Our analytical work based on the layer-peeled model indicates that the feature learning with a fixed ETF classifier naturally leads to the neural collapse state even when the dataset is imbalanced among classes. We further show that in this case the cross entropy (CE) loss is not necessary and can be replaced by a simple squared loss that shares the same global optimality but enjoys a better convergence property. Our experimental results show that our method is able to bring significant improvements with faster convergence on multiple imbalanced datasets.

## Dataset Factorization for Condensation

- Songhua Liu · Kai Wang · Xingyi Yang · Jingwen Ye · Xinchao Wang
- abstract@[open-review](#): In this paper, we study dataset distillation (DD), from a novel perspective and introduce a \emph{dataset factorization} approach, termed \emph{HaBa}, which is a plug-and-play strategy portable to any existing DD baseline. Unlike conventional DD approaches that aim to produce distilled and representative samples, \emph{HaBa} explores decomposing a dataset into two components: data \emph{Ha}llucination networks and \emph{Ba}ses, where the latter is fed into the former to reconstruct image samples. The flexible combinations between bases and hallucination networks, therefore, equip the distilled data with exponential informativeness gain, which largely increase the representation capability of distilled datasets. To guide the compression results to extract more discriminative information, we further introduce a pair of adversarial contrastive constraints on the resultant hallucination networks and bases, which increase the diversity of generated images and inject more discriminant information into the factorization. Extensive comparisons and experiments demonstrate that our method can yield significant improvement on downstream classification tasks compared with previous state of the arts, while reducing the total number of compressed parameters by up to 65\%. Moreover, distilled datasets by our approach also achieve \textasciitilde10\% higher accuracy than baseline methods in cross-architecture generalization.

## Precise Learning Curves and Higher-Order Scalings for Dot-product Kernel Regression

- Lechao Xiao · Jeffrey Pennington · Theodor Misiakiewicz · Hong Hu · Yue Lu
- abstract@[open-review](#): As modern machine learning models continue to advance the computational frontier, it has become increasingly important to develop precise estimates for expected performance improvements under different model and data scaling regimes. Currently, theoretical understanding of the learning curves that characterize how the prediction error depends on the number of samples is restricted to either large-sample asymptotics ( $m \rightarrow \infty$ ) or, for certain simple data distributions, to the high-dimensional asymptotics in which the number of samples scales linearly with the dimension ( $m \propto d$ ). There is a wide gulf between these two regimes, including all higher-order scaling relations  $m \propto d^r$ , which are the subject of the present paper. We focus on the problem of kernel ridge regression for dot-product kernels and present precise formulas for the mean of the test error, bias, and variance, for data drawn uniformly from the sphere with isotropic random labels in the  $r$ -th-order asymptotic scaling regime  $m \rightarrow \infty$  with  $m/d^r$  held constant. We observe a peak in the learning curve whenever  $m \approx d^r/r!$  for any integer  $r$ , leading to multiple sample-wise descent and nontrivial behavior at multiple scales.

## Regret Bounds for Information-Directed Reinforcement Learning

- Botao Hao · Tor Lattimore
- abstract@[open-review](#): Information-directed sampling (IDS) has revealed its potential as a data-efficient algorithm for reinforcement learning (RL). However, theoretical understanding of IDS for Markov Decision Processes (MDPs) is still limited. We develop novel information-theoretic tools to bound the information ratio and cumulative information gain about the learning target. Our theoretical results shed light on the importance of choosing the learning target such that the practitioners can balance the computation and regret bounds. As a consequence, we derive prior-free Bayesian regret bounds for vanilla-IDS which learns the whole environment under tabular finite-horizon MDPs. In addition, we propose a computationally-efficient regularized-IDS that maximizes an additive form rather than the ratio form and show that it enjoys the same regret bound as vanilla-IDS. With the aid of rate-distortion theory, we improve the regret bound by learning a surrogate, less informative environment. Furthermore, we extend our analysis to linear MDPs and prove similar regret bounds for Thompson sampling as a by-product.

## Single-phase deep learning in cortico-cortical networks

- Will Greedy · Heng Wei Zhu · Joseph Pemberton · Jack Mellor · Rui Ponte Costa
- abstract@[open-review](#): The error-backpropagation (backprop) algorithm remains the most common solution to the credit assignment problem in artificial neural networks. In neuroscience, it is unclear whether the brain could adopt a similar strategy to correctly modify its synapses. Recent models have attempted to bridge this gap while being consistent with a range of experimental observations. However, these models are either unable to effectively backpropagate error signals across multiple layers or require a multi-phase learning process, neither of which are reminiscent of learning in the brain. Here, we introduce a new model, bursting cortico-cortical networks (BurstCCN), which solves these issues by integrating known properties of cortical networks namely bursting activity, short-term plasticity (STP) and dendrite-targeting interneurons. BurstCCN relies on burst multiplexing via connection-type-specific STP to propagate backprop-like error signals within deep cortical networks. These error signals are encoded at distal dendrites and induce burst-dependent plasticity as a result of changes to dendritic excitability from excitatory-inhibitory top-down inputs. First, we demonstrate that our model can effectively backpropagate errors through multiple layers using a single-phase learning process. Next, we show both empirically and analytically that learning in our model approximates backprop-derived gradients. Finally, we demonstrate that our model is capable of learning complex image classification tasks (MNIST and CIFAR-10). Overall, our results suggest that cortical features across sub-cellular, cellular, microcircuit and systems levels jointly underlie single-phase efficient deep learning in the brain.

## Explicit Tradeoffs between Adversarial and Natural Distributional Robustness

- Mazda Moayeri · Kiarash Banihashem · Soheil Feizi
- abstract@[open-review](#): Several existing works study either adversarial or natural distributional robustness of deep neural networks separately. In practice, however, models need to enjoy both types of robustness to ensure reliability. In this work, we bridge this gap and show that in fact, {\it explicit tradeoffs} exist between adversarial and natural distributional robustness. We first consider a simple linear regression setting on Gaussian data with disjoint sets of \emph{core} and \emph{spurious} features. In this setting, through theoretical and empirical analysis, we show that (i) adversarial training with  $\ell_1$  and  $\ell_2$  norms increases the model reliance on spurious features; (ii) For  $\ell_\infty$  adversarial training, spurious reliance only occurs when the scale of the spurious features is larger than that of the core features; (iii) adversarial training can have {\it an unintended consequence} in reducing distributional robustness, specifically when spurious correlations are changed in the new test domain. Next, we present extensive empirical evidence, using a test suite of twenty adversarially trained models evaluated on five benchmark datasets (ObjectNet, RIVAL10, Salient ImageNet-1M, ImageNet-9, Waterbirds), that adversarially trained classifiers rely on backgrounds more than their standardly trained counterparts, validating our theoretical results. We also show that spurious correlations in training data (when preserved in the test domain) can {\it improve} adversarial robustness, revealing that previous claims that adversarial vulnerability is rooted in spurious correlations are incomplete.

## Unsupervised Representation Learning from Pre-trained Diffusion Probabilistic Models

- Zijian Zhang · Zhou Zhao · Zhijie Lin
- abstract@[open-review](#): Diffusion Probabilistic Models (DPMs) have shown a powerful capacity of generating high-quality image samples. Recently, diffusion autoencoders (Diff-AE) explore DPMs for representation learning via autoencoding and succeed in various downstream tasks. Their key idea is to jointly train an encoder for discovering meaningful representations from images and a conditional DPM as the decoder for image reconstruction. Considering that training DPMs will take a long time and there have already been many pre-trained unconditional DPMs, we aim to adapt these pre-trained models for representation learning also via autoencoding, but with less training times and better performance than Diff-AE. We find that the key factor that pre-trained DPMs cannot reconstruct image from latent variable is the information loss of forward diffusion process, which causes a gap between the predicted posterior mean by pre-trained DPMs and the true posterior mean. Inspired by the classifier-guided sampling method, we employ an encoder to learn meaningful representations from images and a gradient estimator to directly model the mean shift according to the learned representations to fill the posterior mean gap for image reconstruction. By further leveraging the knowledges of pre-trained DPMs and redesigning the weighting scheme

of training objective, our method can learn richer representations from images more efficiently. Extensive experiments show that our method outperforms Diff-AE and enables some added tasks and features. We will make the code publicly available shortly.

## [MCL-GAN: Generative Adversarial Networks with Multiple Specialized Discriminators](#)

- Jinyoung Choi · Bohyung Han
- abstract@[open-review](#): We propose a generative adversarial network with multiple discriminators, which collaborate to represent a real dataset more effectively. This approach facilitates learning a generator consistent with the underlying data distribution based on real images and thus mitigates the chronic mode collapse problem. From the inspiration of multiple choice learning, we guide each discriminator to have expertise in the subset of the entire data and allow the generator to find reasonable correspondences between the latent and real data spaces automatically without the extra supervision for training examples. Despite the use of multiple discriminators, the backbone networks are shared across the discriminators and the increase of training cost is marginal. We demonstrate the effectiveness of our algorithm using multiple evaluation metrics in the standard datasets for diverse tasks.

## [Remember the Past: Distilling Datasets into Addressable Memories for Neural Networks](#)

- Zhiwei Deng · Olga Russakovsky
- abstract@[open-review](#): We propose an algorithm that compresses the critical information of a large dataset into compact addressable memories. These memories can then be recalled to quickly re-train a neural network and recover the performance (instead of storing and re-training on the full original dataset). Building upon the dataset distillation framework, we make a key observation that a shared common representation allows for more efficient and effective distillation. Concretely, we learn a set of bases (aka ``memories'') which are shared between classes and combined through learned flexible addressing functions to generate a diverse set of training examples. This leads to several benefits: 1) the size of compressed data does not necessarily grow linearly with the number of classes; 2) an overall higher compression rate with more effective distillation is achieved; and 3) more generalized queries are allowed beyond recalling the original classes. We demonstrate state-of-the-art results on the dataset distillation task across five benchmarks, including up to 16.5% and 9.7% accuracy improvement when distilling CIFAR10 and CIFAR100 respectively. We then leverage our framework to perform continual learning, achieving state-of-the-art results on four benchmarks, with 23.2% accuracy improvement on MANY.

## [Relational Language-Image Pre-training for Human-Object Interaction Detection](#)

- Hangjie Yuan · Jianwen Jiang · Samuel Albanie · Tao Feng · Ziyuan Huang · Dong Ni · Mingqian Tang
- abstract@[open-review](#): The task of Human-Object Interaction (HOI) detection targets fine-grained visual parsing of humans interacting with their environment, enabling a broad range of applications. Prior work has demonstrated the benefits of effective architecture design and integration of relevant cues for more accurate HOI detection. However, the design of an appropriate pre-training strategy for this task remains underexplored by existing approaches. To address this gap, we propose \$\textit{Relational Language-Image Pre-training}\$ (RLIP), a strategy for contrastive pre-training that leverages both entity and relation descriptions. To make effective use of such pre-training, we make three technical contributions: (1) a new \$textbf{Par}\$\_allel entity detection and \$textbf{Se}\$\_quential relation inference (ParSe) architecture that enables the use of both entity and relation descriptions during holistically optimized pre-training; (2) a synthetic data generation framework, Label Sequence Extension, that expands the scale of language data available within each minibatch; (3) ambiguity-suppression mechanisms, Relation Quality Labels and Relation Pseudo-Labels, to mitigate the influence of ambiguous/noisy samples in the pre-training data. Through extensive experiments, we demonstrate the benefits of these contributions, collectively termed RLIP-ParSe, for improved zero-shot, few-shot and fine-tuning HOI detection performance as well as increased robustness to learning from noisy annotations.

## [BEER: Fast \\$O\(1/T\)\\$ Rate for Decentralized Nonconvex Optimization with Communication Compression](#)

- Haoyu Zhao · Boyue Li · Zhize Li · Peter Richtarik · Yuejie Chi
- abstract@[open-review](#): Communication efficiency has been widely recognized as the bottleneck for large-scale decentralized machine learning applications in multi-agent or federated environments. To tackle the communication bottleneck, there have been many efforts to design communication-compressed algorithms for decentralized nonconvex optimization, where the clients are only allowed to communicate a small amount of quantized information (aka bits) with their neighbors over a predefined graph topology. Despite significant efforts, the state-of-the-art algorithm in the nonconvex setting still suffers from a slower rate of convergence \$O((G/T)^{2/3})\$ compared with their uncompressed counterpart, where \$G\$ measures the data heterogeneity across different clients, and \$T\$ is the number of communication rounds. This paper proposes BEER, which adopts communication compression with gradient tracking, and shows it converges at a faster rate of \$O(1/T)\$. This significantly improves over the state-of-the-art rate, by matching the rate without compression even under arbitrary data heterogeneity. Numerical experiments are also provided to corroborate our theory and confirm the practical superiority of beer in the data heterogeneous regime.

## [Trap and Replace: Defending Backdoor Attacks by Trapping Them into an Easy-to-Replace Subnetwork](#)

- Haotao Wang · Junyuan Hong · Aston Zhang · Jiayu Zhou · Zhangyang Wang
- abstract@[open-review](#): Deep neural networks (DNNs) are vulnerable to backdoor attacks. Previous works have shown it extremely challenging to unlearn the undesired backdoor behavior from the network, since the entire network can be affected by the backdoor samples. In this paper, we propose a brand-new backdoor defense strategy, which makes it much easier to remove the harmful influence of backdoor samples from the model. Our defense strategy, \textit{Trap and Replace}, consists of two stages. In the first stage, we bait and trap the backdoors in a small and easy-to-replace subnetwork. Specifically, we add an auxiliary image reconstruction head on top of the stem network shared with a light-weighted classification head. The intuition is that the auxiliary image reconstruction task encourages the stem network to keep sufficient low-level visual features that are hard to learn but semantically correct, instead of overfitting to the easy-to-learn but semantically incorrect backdoor correlations. As a result, when trained on backdoored datasets, the backdoors are easily baited towards the unprotected classification head, since it is much more vulnerable than the shared stem, leaving the stem network hardly poisoned. In the second stage, we replace the poisoned light-weighted classification head with an untainted one, by re-training it from scratch only on a small holdout dataset with clean samples, while fixing the stem network. As a result, both the stem and the classification head in the final network are hardly affected by backdoor training samples. We evaluate our method against ten different backdoor attacks. Our method outperforms previous state-of-the-art methods by up to 20.57%, 9.80%, and 13.72% attack success rate and on-average 3.14%, 1.80%, and 1.21% clean classification accuracy on CIFAR10, GTSRB, and ImageNet-12, respectively. Source code and pre-trained models will be released.

## [HSurf-Net: Normal Estimation for 3D Point Clouds by Learning Hyper Surfaces](#)

- Qing Li · Yu-Shen Liu · Jin-San Cheng · Cheng Wang · Yi Fang · Zhizhong Han
- abstract@[open-review](#): We propose a novel normal estimation method called HSurf-Net, which can accurately predict normals from point clouds with noise and density variations. Previous methods focus on learning point weights to fit neighborhoods into a geometric surface approximated by a polynomial function of a predefined order, based on which normals are estimated. However, fitting surfaces explicitly from raw point clouds suffers from overfitting or underfitting issues caused by inappropriate polynomial orders and outliers, which significantly limits the performance of the existing methods. To address these issues, we introduce hyper surface fitting to implicitly learn hyper surfaces, which are represented by multi-layer perceptron (MLP) layers that take point features and output surface patterns in a high dimensional feature space. We employ a novel space transformation module, which consists of a sequence of local aggregation layers and global shift layers, to reliably build the feature space, and a relative position encoding module to effectively convert the point clouds into that feature space. Our model learns hyper surfaces from the noise-less features and directly predicts normal

vectors. We jointly optimize the MLP weights and module parameters in a data-driven manner to make the model adaptively find the most suitable surface pattern for various points. Experimental results show that our HSurf-Net achieves state-of-the-art (SOTA) performance on the synthetic shape dataset, the real-world indoor and outdoor scene datasets.

## [Inception Transformer](#)

- Chenyang Si · Weihao Yu · Pan Zhou · Yichen Zhou · Xinchao Wang · Zhongwen Xu
- abstract@[open-review](#): Recent studies show that transformer has strong capability of building long-range dependencies, yet is incompetent in capturing high frequencies that predominantly convey local information. To tackle this issue, we present a novel and general-purpose \$textit{Inception Transformer}\$, or \$textit{iFormer}\$ for short, that effectively learns comprehensive features with both high- and low-frequency information in visual data. Specifically, we design an Inception mixer to explicitly graft the advantages of convolution and max-pooling for capturing the high-frequency information to transformers. Different from recent hybrid frameworks, the Inception mixer brings greater efficiency through a channel splitting mechanism to adopt parallel convolution/max-pooling path and self-attention path as high- and low-frequency mixers, while having the flexibility to model discriminative information scattered within a wide frequency range. Considering that bottom layers play more roles in capturing high-frequency details while top layers more in modeling low-frequency global information, we further introduce a frequency ramp structure, i.e., gradually decreasing the dimensions fed to the high-frequency mixer and increasing those to the low-frequency mixer, which can effectively trade-off high- and low-frequency components across different layers. We benchmark the iFormer on a series of vision tasks, and showcase that it achieves impressive performance on image classification, COCO detection and ADE20K segmentation. For example, our iFormer-S hits the top-1 accuracy of 83.4% on ImageNet-1K, much higher than DeiT-S by 3.6%, and even slightly better than much bigger model Swin-B (83.3%) with only 1/4 parameters and 1/3 FLOPs. Code and models will be released.

## [DOMINO: Decomposed Mutual Information Optimization for Generalized Context in Meta-Reinforcement Learning](#)

- Yao Mu · Yuzheng Zhuang · Fei Ni · Bin Wang · Jianyu Chen · Jianye Hao · Ping Luo
- abstract@[open-review](#): Adapting to the changes in transition dynamics is essential in robotic applications. By learning a conditional policy with a compact context, context-aware meta-reinforcement learning provides a flexible way to adjust behavior according to dynamics changes. However, in real-world applications, the agent may encounter complex dynamics changes. Multiple confounders can influence the transition dynamics, making it challenging to infer accurate context for decision-making. This paper addresses such a challenge by decomposed mutual information optimization (DOMINO) for context learning, which explicitly learns a disentangled context to maximize the mutual information between the context and historical trajectories while minimizing the state transition prediction error. Our theoretical analysis shows that DOMINO can overcome the underestimation of the mutual information caused by multi-confounded challenges via learning disentangled context and reduce the demand for the number of samples collected in various environments. Extensive experiments show that the context learned by DOMINO benefits both model-based and model-free reinforcement learning algorithms for dynamics generalization in terms of sample efficiency and performance in unseen environments.

## [Align then Fusion: Generalized Large-scale Multi-view Clustering with Anchor Matching Correspondences](#)

- Siwei Wang · Xinwang Liu · Suyuan Liu · Jiaqi Jin · Wenzuan Tu · Xinzhang Zhu · En Zhu
- abstract@[open-review](#): Multi-view anchor graph clustering selects representative anchors to avoid full pair-wise similarities and therefore reduce the complexity of graph methods. Although widely applied in large-scale applications, existing approaches do not pay sufficient attention to establishing correct correspondences between the anchor sets across views. To be specific, anchor graphs obtained from different views are not aligned column-wisely. Such an Anchor-Unaligned Problem (AUP) would cause inaccurate graph fusion and degrade the clustering performance. Under multi-view scenarios, generating correct correspondences could be extremely difficult since anchors are not consistent in feature dimensions. To solve this challenging issue, we propose the first study of a generalized and flexible anchor graph fusion framework termed Fast Multi-View Anchor-Correspondence Clustering (FMVACC). Specifically, we show how to find anchor correspondence with both feature and structure information, after which anchor graph fusion is performed column-wisely. Moreover, we theoretically show the connection between FMVACC and existing multi-view late fusion and partial view-aligned clustering which further demonstrates our generality. Extensive experiments on seven benchmark datasets demonstrate the effectiveness and efficiency of our proposed method. Moreover, the proposed alignment module also shows significant performance improvement applying to existing multi-view anchor graph competitors indicating the importance of anchor alignment.

## [Differentiable Analog Quantum Computing for Optimization and Control](#)

- Jiaqi Leng · Yuxiang Peng · Yi-Ling Qiao · Ming Lin · Xiaodi Wu
- abstract@[open-review](#): We formulate the first differentiable analog quantum computing framework with specific parameterization design at the analog signal (pulse) level to better exploit near-term quantum devices via variational methods. We further propose a scalable approach to estimate the gradients of quantum dynamics using a forward pass with Monte Carlo sampling, which leads to a quantum stochastic gradient descent algorithm for scalable gradient-based training in our framework. Applying our framework to quantum optimization and control, we observe a significant advantage of differentiable analog quantum computing against SOTAs based on parameterized digital quantum circuits by \$\{\backslash em\$ orders of magnitude\$\}.

## [Long-Form Video-Language Pre-Training with Multimodal Temporal Contrastive Learning](#)

- Yuchong Sun · Bei Liu · Hongwei Xue · Ruihua Song · Huan Yang · Jianlong Fu
- abstract@[open-review](#): Large-scale video-language pre-training has shown significant improvement on video-language understanding tasks. Previous studies of video-language pre-training mainly focus on short-form videos (i.e., within 30 seconds) and sentences, leaving long-form video-language pre-training rarely explored. Directly learning representation from long-form videos and language is challenging due to the difficulty of modeling long-range relationship and heavy computation burden caused by more frames. In this paper, we introduce a Long-Form Video-Language Pre-training (LF-VLP) model, and train it on a large-scale long-form video and paragraph dataset constructed from an existing public dataset. To effectively capture the rich temporal dynamics and to better align video and language in an efficient end-to-end manner, we introduce two novel designs in our LF-VLP model. We first propose a multimodal temporal contrastive loss (MTC) to learn the temporal relation across different modalities by encouraging fine-grained alignment between long-form videos and paragraphs. Second, we propose a hierarchical temporal window attention (HTWA) mechanism to effectively capture long-range dependency while reducing computational cost in Transformer. We fine-tune the pre-trained LF-VLP model on seven downstream long-form video-language understanding tasks of paragraph-to-video retrieval or long-form video QA, and achieve the new state-of-the-art performances. Specifically, our model achieves 16.1% relative improvement on ActivityNet paragraph-to-video retrieval task and 2.4% on How2QA task, respectively.

## [Fine-Grained Semantically Aligned Vision-Language Pre-Training](#)

- Juncheng Li · XIN HE · Longhui Wei · Long Qian · Linchao Zhu · Lingxi Xie · Yueling Zhuang · Qi Tian · Siliang Tang
- abstract@[open-review](#): Large-scale vision-language pre-training has shown impressive advances in a wide range of downstream tasks. Existing methods mainly model the cross-modal alignment by the similarity of the global representations of images and text, or advanced cross-modal attention upon image and text features. However, they fail to explicitly learn the fine-grained semantic alignment between visual regions and textual phrases, as only global image-text alignment information is available. In this paper, we introduce LOUPE, a fine-grained semantically aligned visiOn-langUage PrE-training framework, which learns fine-grained semantic alignment from the novel perspective of game-theoretic interactions. To efficiently estimate the game-theoretic interactions, we further propose an uncertainty-aware neural Shapley interaction learning module. Experiments show that LOUPE achieves state-

of-the-art performance on a variety of vision-language tasks. Without any object-level human annotations and fine-tuning, LOUPE achieves competitive performance on object detection and visual grounding. More importantly, LOUPE opens a new promising direction of learning fine-grained semantics from large-scale raw image-text pairs.

## Understanding Neural Architecture Search: Convergence and Generalization

- Zhenyu Zhu · Fanghui Liu · Grigoris Chrysos · Volkan Cevher
- abstract@[open-review](#): Neural Architecture Search (NAS) has fostered the automatic discovery of neural architectures, which achieve state-of-the-art accuracy in image recognition. Despite the progress achieved with NAS, so far there is little attention to theoretical guarantees on NAS. In this work, we study the generalization properties of NAS under a unifying framework enabling (deep) layer skip connection search and activation function search. To this end, we derive the lower (and upper) bounds of the minimum eigenvalue of Neural Tangent Kernel under the (in)finite width regime from a search space including mixed activation functions, fully connected, and residual neural networks. Our analysis is non-trivial due to the coupling of various architectures and activation functions under the unifying framework. Then, we leverage the eigenvalue bounds to establish generalization error bounds of NAS in the stochastic gradient descent training. Importantly, we theoretically and experimentally show how the derived results can guide NAS to select the top-performing architectures, even in the case without training, leading to a training-free algorithm based on our theory. Accordingly, our numerical validation shed light on the design of computationally efficient methods for NAS.

## Efficient Adversarial Training without Attacking: Worst-Case-Aware Robust Reinforcement Learning

- Yongyuan Liang · Yanchao Sun · Ruijie Zheng · Furong Huang
- abstract@[open-review](#): Recent studies reveal that a well-trained deep reinforcement learning (RL) policy can be particularly vulnerable to adversarial perturbations on input observations. Therefore, it is crucial to train RL agents that are robust against any attacks with a bounded budget. Existing robust training methods in deep RL either treat correlated steps separately, ignoring the robustness of long-term reward, or train the agents and RL-based attacker together, doubling the computational burden and sample complexity of the training process. In this work, we propose a strong and efficient robust training framework for RL, named Worst-case-aware Robust RL (WocaR-RL), that directly estimates and optimizes the worst-case reward of a policy under bounded  $\ell_p$  attacks without requiring extra samples for learning an attacker. Experiments on multiple environments show that WocaR-RL achieves state-of-the-art performance under various strong attacks, and obtains significantly higher training efficiency than prior state-of-the-art robust training methods.

## On the Epistemic Limits of Personalized Prediction

- Lucas Monteiro Paes · Carol Long · Berk Ustun · Flavio Calmon
- abstract@[open-review](#): Predictive models often include group attributes that encode personal characteristics like sex, blood type, or HIV status. Personalized models must ensure fair use -- i.e., groups who provide personal data should expect to receive a tailored improvement in accuracy compared to a non-personalized model. In this paper, we derive conditions under which one can detect fair use violations in predictive models and characterize when estimating fair use is impossible. We propose a metric to evaluate the worst-case accuracy gain across groups called the benefit of personalization (BoP). We obtain bounds on the error probability for testing if the BoP is above a target threshold given finite samples. Remarkably, our bounds provide an information-theoretic limit on the number of group attributes that a model can use to allow for verification -- beyond this limit, it is impossible to reliably detect if personalization harms or benefits all groups. We also derive statistical limits for the minimax mean-square error of estimating the BoP. Our results show that there is no way to reliably determine if a personalized model with  $k \geq 19$  attributes benefits every group that provides personal data, even when we are given a dataset with  $N = 8 \times 10^9$  samples (i.e., one observation for each person in the world).

## Black-box pseudodata variational inference

- Dionysis Manousakas · Hippolyt Ritter · Theofanis Karaletsos
- abstract@[open-review](#): In the context of Bayesian inference, recent advances in coresets methods have shown that careful selection of representative datapoints can replace massive volumes of data, preserving the relevant statistical information and significantly accelerating subsequent downstream tasks. Existing variational coresets constructions rely on either selecting subsets of the observed datapoints, or jointly performing approximate inference and optimizing pseudodata in the observed space akin to inducing points methods in Gaussian Processes. So far, both approaches are limited by complexities in evaluating their objectives for general purpose models, and require generating samples from a typically intractable posterior over the coresets throughout inference and testing. In this work, we present a black-box variational inference algorithm for coresets that overcomes these constraints and enables principled application of variational coresets to intractable models, such as Bayesian neural networks. We apply our techniques to supervised learning problems, and compare it with existing approaches in the literature for data summarization and inference.

## The Phenomenon of Policy Churn

- Tom Schaul · Andre Barreto · John Quan · Georg Ostrovski
- abstract@[open-review](#): We identify and study the phenomenon of policy churn, that is, the rapid change of the greedy policy in value-based reinforcement learning. Policy churn operates at a surprisingly rapid pace, changing the greedy action in a large fraction of states within a handful of learning updates (in a typical deep RL set-up such as DQN on Atari). We characterise the phenomenon empirically, verifying that it is not limited to specific algorithm or environment properties. A number of ablations help whittle down the plausible explanations on why churn occurs to just a handful, all related to deep learning. Finally, we hypothesise that policy churn is a beneficial but overlooked form of implicit exploration that casts  $\epsilon$ -greedy exploration in a fresh light, namely that  $\epsilon$ -noise plays a much smaller role than expected.

## Extreme Compression for Pre-trained Transformers Made Simple and Efficient

- Xiaoxia Wu · Zhewei Yao · Minjia Zhang · Conglong Li · Yuxiong He
- abstract@[open-review](#): Extreme compression, particularly ultra-low bit precision (binary/ternary) quantization, has been proposed to fit large NLP models on resource-constraint devices. However, to preserve the accuracy for such aggressive compression schemes, cutting-edge methods usually introduce complicated compression pipelines, e.g., multi-stage expensive knowledge distillation with extensive hyperparameter tuning. Also, they oftentimes focus less on smaller transformer models that have already been heavily compressed via knowledge distillation and lack a systematic study to show the effectiveness of their methods. In this paper, we perform a very comprehensive systematic study to measure the impact of many key hyperparameters and training strategies from previous. As a result, we find out that previous baselines for ultra-low bit precision quantization are significantly under-trained. Based on our study, we propose a simple yet effective compression pipeline for extreme compression. Our simplified pipeline demonstrates that (1) we can skip the pre-training knowledge distillation to obtain a 5-layer BERT while achieving better performance than previous state-of-the-art methods, like TinyBERT; (2) extreme quantization plus layer reduction is able to reduce the model size by 50x, resulting in new state-of-the-art results on GLUE tasks.

## Sample Constrained Treatment Effect Estimation

- Raghavendra Addanki · David Arbour · Tung Mai · Cameron Musco · Anup Rao

- abstract@[open-review](#): Treatment effect estimation is a fundamental problem in causal inference. We focus on designing efficient randomized controlled trials, to accurately estimate the effect of some treatment on a population of  $n$  individuals. In particular, we study sample-constrained treatment effect estimation, where we must select a subset of  $n$  individuals from the population to experiment on. This subset must be further partitioned into treatment and control groups. Algorithms for partitioning the full population into treatment and control groups, or for choosing a single representative subset, have been well-studied. The key challenge in our setting is jointing choosing a representative subset and a partition for that set. We focus on both individual and average treatment effect estimation, under a linear effects model. We give provably efficient experimental designs and corresponding estimators, by identifying connections to discrepancy minimization and leverage-score-based sampling used in randomized numerical linear algebra. Our theoretical results obtain a smooth transition to known guarantees when  $n$  equals the population size. We also empirically demonstrate the performance of our algorithms.

### [Revisit last-iterate convergence of mSGD under milder requirement on step size](#)

- ruinan Jin · Xingkang He · Lang Chen · Difei Cheng · Vijay Gupta
- abstract@[open-review](#): Understanding convergence of SGD-based optimization algorithms can help deal with enormous machine learning problems. To ensure last-iterate convergence of SGD and momentum-based SGD (mSGD), the existing studies usually constrain the step size  $\epsilon_n$  to decay as  $\sum_{n=1}^{+\infty} \epsilon_n^2 < +\infty$ , which however is rather conservative and may lead to slow convergence in the early stage of the iteration. In this paper, we relax this requirement by studying an alternate step size for the mSGD. First, we relax the requirement of the decay on step size to  $\sum_{n=1}^{+\infty} \epsilon_n^{2+\eta} < +\infty$  ( $0 < \eta < 1/2$ ). This implies that a larger step size, such as  $\epsilon_n = \frac{1}{\sqrt{n}}$  can be utilized for accelerating the mSGD in the early stage. Under this new step size and some common conditions, we prove that the gradient norm of mSGD for non-convex loss functions asymptotically decays to zero. In addition, we show that this step size can indeed help make the convergence into a neighborhood of the stationary points quicker in the early stage. In addition, we establish the convergence of mSGD under a constant step size  $\epsilon_n \equiv \epsilon > 0$  by removing the common requirement in the literature on the strong convexity of the loss function. Some experiments are given to illustrate the developed results.

### [GAL: Gradient Assisted Learning for Decentralized Multi-Organization Collaborations](#)

- Enmao Diao · Jie Ding · Vahid Tarokh
- abstract@[open-review](#): Collaborations among multiple organizations, such as financial institutions, medical centers, and retail markets in decentralized settings are crucial to providing improved service and performance. However, the underlying organizations may have little interest in sharing their local data, models, and objective functions. These requirements have created new challenges for multi-organization collaboration. In this work, we propose Gradient Assisted Learning (GAL), a new method for multiple organizations to assist each other in supervised learning tasks without sharing local data, models, and objective functions. In this framework, all participants collaboratively optimize the aggregation of local loss functions, and each participant autonomously builds its own model by iteratively fitting the gradients of the overarching objective function. We also provide asymptotic convergence analysis and practical case studies of GAL. Experimental studies demonstrate that GAL can achieve performance close to centralized learning when all data, models, and objective functions are fully disclosed.

### [Towards Understanding the Mixture-of-Experts Layer in Deep Learning](#)

- Zixiang Chen · Yihe Deng · Yue Wu · Quanquan Gu · Yuanzhi Li
- abstract@[open-review](#): The Mixture-of-Experts (MoE) layer, a sparsely-activated model controlled by a router, has achieved great success in deep learning. However, the understanding of such architecture remains elusive. In this paper, we formally study how the MoE layer improves the performance of neural network learning and why the mixture model will not collapse into a single model. Our empirical results suggest that the cluster structure of the underlying problem and the non-linearity of the expert are pivotal to the success of MoE. This motivates us to consider a challenging classification problem with intrinsic cluster structures. Theoretically, we proved that this problem is hard to solve by a single expert such as a two-layer convolutional neural network (CNN). Yet with the MoE layer with each expert being a two-layer CNN, the problem can be solved successfully. In particular, our theory shows that the router can learn the cluster-center features, which helps divide the input complex problem into simpler classification sub-problems that individual experts can conquer. To our knowledge, this is the first theoretical result toward formally understanding the mechanism of the MoE layer for deep learning.

### [Outlier-Robust Sparse Estimation via Non-Convex Optimization](#)

- Yu Cheng · Ilias Diakonikolas · Rong Ge · Shivam Gupta · Daniel Kane · Mahdi Soltanolkotabi
- abstract@[open-review](#): We explore the connection between outlier-robust high-dimensional statistics and non-convex optimization in the presence of sparsity constraints, with a focus on the fundamental tasks of robust sparse mean estimation and robust sparse PCA. We develop novel and simple optimization formulations for these problems such that any approximate stationary point of the associated optimization problem yields a near-optimal solution for the underlying robust estimation task. As a corollary, we obtain that any first-order method that efficiently converges to stationarity yields an efficient algorithm for these tasks. The obtained algorithms are simple, practical, and succeed under broader distributional assumptions compared to prior work.

### [Expected Frequency Matrices of Elections: Computation, Geometry, and Preference Learning](#)

- Niclas Boehmer · Robert Bredereck · Edith Elkind · Piotr Faliszewski · Stanisław Szufa
- abstract@[open-review](#): We use the "map of elections" approach of Szufa et al. (AAMAS 2020) to analyze several well-known vote distributions. For each of them, we give an explicit formula or an efficient algorithm for computing its frequency matrix, which captures the probability that a given candidate appears in a given position in a sampled vote. We use these matrices to draw the "skeleton map" of distributions, evaluate its robustness, and analyze its properties. We further develop a general and unified framework for learning the distribution of real-world preferences using the frequency matrices of established vote distributions.

### [Counterfactual harm](#)

- Jonathan Richens · Rory Beard · Daniel H. Thompson
- abstract@[open-review](#): To act safely and ethically in the real world, agents must be able to reason about harm and avoid harmful actions. However, to date there is no definition of harm that can be incorporated into machine learning algorithms. In this paper we propose the first statistical definition of harm based on the predominant `counterfactual comparative account'. We show that any factual definition of harm must violate basic intuitions in certain scenarios, and show that standard machine learning algorithms that cannot perform counterfactual reasoning are guaranteed to pursue harmful policies following certain distributional shifts. To resolve this we derive a family of counterfactual objective functions that robustly mitigate for harm. We demonstrate our framework on a real-world problem of identifying optimal drug doses, using a dose-response model learned from a meta-analysis for randomized control trial data.

### [Video-based Human-Object Interaction Detection from Tubelet Tokens](#)

- Danyang Tu · Wei Sun · Xiongkuo Min · Guangtao Zhai · Wei Shen
- abstract@[open-review](#): We present a novel vision Transformer, named TUTOR, which is able to learn tubelet tokens, served as highly-abstacted spatial-temporal representations, for video-based human-object interaction (V-HOI) detection. The tubelet tokens structurize videos by agglomerating and linking semantically-related patch tokens along spatial and temporal domains, which enjoy two benefits: 1) Compactness: each token is learned by a selective attention mechanism to reduce redundant dependencies from others; 2) Expressiveness: each token is enabled to align with a semantic instance, i.e., an object or a human, thanks to agglomeration and linking. The effectiveness and efficiency of TUTOR are verified by extensive experiments. Results show our method outperforms existing works by large margins, with a relative mAP gain of \$16.14\%\$ on VidHOI and a 2 points gain on CAD-120 as well as a \$4\times\$ speedup.

## Chaotic Regularization and Heavy-Tailed Limits for Deterministic Gradient Descent

- Soon Hoe Lim · Yijun Wan · Umut Simsekli
- abstract@[open-review](#): Recent studies have shown that gradient descent (GD) can achieve improved generalization when its dynamics exhibits a chaotic behavior. However, to obtain the desired effect, the step-size should be chosen sufficiently large, a task which is problem dependent and can be difficult in practice. In this study, we incorporate a chaotic component to GD in a controlled manner, and introduce \text{multiscale perturbed GD} (MPGD), a novel optimization framework where the GD recursion is augmented with chaotic perturbations that evolve via an independent dynamical system. We analyze MPGD from three different angles: (i) By building up on recent advances in rough paths theory, we show that, under appropriate assumptions, as the step-size decreases, the MPGD recursion converges weakly to a stochastic differential equation (SDE) driven by a heavy-tailed Lévy-stable process. (ii) By making connections to recently developed generalization bounds for heavy-tailed processes, we derive a generalization bound for the limiting SDE and relate the worst-case generalization error over the trajectories of the process to the parameters of MPGD. (iii) We analyze the implicit regularization effect brought by the dynamical regularization and show that, in the weak perturbation regime, MPGD introduces terms that penalize the Hessian of the loss function. Empirical results are provided to demonstrate the advantages of MPGD.

## WaveBound: Dynamically Bounding Error for Stable Time Series Forecasting

- Youngin Cho · Daejin Kim · DONGMIN KIM · MOHAMMAD AZAM KHAN · Jaegul Choo
- abstract@[open-review](#): Time series forecasting becomes a critical task due to its high practicality in real-world applications such as traffic, energy consumption, economics and finance, and disease analysis. Recent deep-learning-based approaches have shown remarkable success in time series forecasting. Nonetheless, due to the dynamics of time series data, deep networks still suffer from unstable training and overfitting. Inconsistent patterns appearing in real-world data lead the model to be biased to a particular pattern, thus limiting the generalization. In this work, we introduce the dynamic error bounds on training loss to address the overfitting issue in time series forecasting. Consequently, we propose a regularization method called WaveBound which estimates the adequate error bounds of training loss for each time step and feature at each iteration. By allowing the model to focus less on unpredictable data, WaveBound stabilizes the training process, hence significantly improving generalization. With the extensive experiments, we show that WaveBound consistently improves the existing models in large margins, including the state-of-the-art model.

## Multi-objective Deep Data Generation with Correlated Property Control

- Shiyu Wang · Xiaojie Guo · Xuanyang Lin · Bo Pan · Yuanqi Du · Yinkai Wang · Yanfang Ye · Ashley Petersen · Austin Leitgeb · Saleh Alkhaila · Kevin Minbile · William M. Wuest · Amarda Shehu · Liang Zhao
- abstract@[open-review](#): Developing deep generative models has been an emerging field due to the ability to model and generate complex data for various purposes, such as image synthesis and molecular design. However, the advance of deep generative models is limited by the challenges to generate objects that possess multiple desired properties because: 1) the existence of complex correlation among real-world properties is common but hard to identify; 2) controlling individual property enforces an implicit partially control of its correlated properties, which is difficult to model; 3) controlling multiple properties under various manners simultaneously is hard and underexplored. We address these challenges by proposing a novel deep generative framework that recovers semantics and correlation of properties through disentangled latent vectors. The correlation is handled via an explainable mask pooling layer, and properties are precisely retained by the generated objects via the mutual dependence between latent vectors and properties. Our generative model preserves properties of interest while handles correlation and conflicts of properties under a multi-objective optimization framework. The experiments demonstrate our model's superior performance in generating objects with desired properties.

## Recursive Reasoning in Minimax Games: A Level $\$k\$$ Gradient Play Method

- Zichu Liu · Lacra Pavel
- abstract@[open-review](#): Despite the success of generative adversarial networks (GANs) in generating visually appealing images, they are notoriously challenging to train. In order to stabilize the learning dynamics in minimax games, we propose a novel recursive reasoning algorithm: Level  $\$k\$$  Gradient Play (Lv. $\$k\$$  GP) algorithm. Our algorithm does not require sophisticated heuristics or second-order information, as do existing algorithms based on predictive updates. We show that as  $k$  increases, Lv. $\$k\$$  GP converges asymptotically towards an accurate estimation of players' future strategy. Moreover, we justify that Lv. $\$infty\$$  GP naturally generalizes a line of provably convergent game dynamics which rely on predictive updates. Furthermore, we provide its local convergence property in nonconvex-nonconcave zero-sum games and global convergence in bilinear and quadratic games. By combining Lv. $\$k\$$  GP with Adam optimizer, our algorithm shows a clear advantage in terms of performance and computational overhead compared to other methods. Using a single Nvidia RTX3090 GPU and 30 times fewer parameters than BigGAN on CIFAR-10, we achieve an FID of 10.17 for unconditional image generation within 24 hours, allowing GAN training on common computational resources to reach state-of-the-art performance.

## Pre-Train Your Loss: Easy Bayesian Transfer Learning with Informative Priors

- Ravid Shwartz-Ziv · Micah Goldblum · Hossein Souri · Sanyam Kapoor · Chen Zhu · Yann LeCun · Andrew Wilson
- abstract@[open-review](#): Deep learning is increasingly moving towards a transfer learning paradigm whereby large ``foundation models'' are fine-tuned on downstream tasks, starting from an initialization learned on the source task. But an initialization contains relatively little information about the source task. Instead, we show that we can learn highly informative posteriors from the source task, through supervised or self-supervised approaches, which then serve as the basis for priors that modify the whole loss surface on the downstream task. This simple modular approach enables significant performance gains and more data-efficient learning on a variety of downstream classification and segmentation tasks, serving as a drop-in replacement for standard pre-training strategies. These highly informative priors also can be saved for future use, similar to pre-trained weights, and stand in contrast to the zero-mean isotropic uninformative priors that are typically used in Bayesian deep learning.

## Representing Spatial Trajectories as Distributions

- Didac Suris Coll-Vinent · Carl Vondrick
- abstract@[open-review](#): We introduce a representation learning framework for spatial trajectories. We represent partial observations of trajectories as probability distributions in a learned latent space, which characterize the uncertainty about unobserved parts of the trajectory. Our framework allows us to obtain samples from a trajectory for any continuous point in time—both interpolating and extrapolating. Our flexible approach supports directly modifying specific attributes of a trajectory, such as its pace, as well as combining different partial observations into single representations. Experiments show our method's superiority over baselines in prediction tasks.

## SparCL: Sparse Continual Learning on the Edge

- Zifeng Wang · Zheng Zhan · Yifan Gong · Geng Yuan · Wei Niu · Tong Jian · Bin Ren · Stratis Ioannidis · Yanzhi Wang · Jennifer Dy
- abstract@[open-review](#): Existing work in continual learning (CL) focuses on mitigating catastrophic forgetting, i.e., model performance deterioration on past tasks when learning a new task. However, the training efficiency of a CL system is under-investigated, which limits the real-world application of CL systems under resource-limited scenarios. In this work, we propose a novel framework called Sparse Continual Learning (SparCL), which is the first study that leverages sparsity to enable cost-effective continual learning on edge devices. SparCL achieves both training acceleration and accuracy preservation through the synergy of three aspects: weight sparsity, data efficiency, and gradient sparsity. Specifically, we propose task-aware dynamic masking (TDM) to learn a sparse network throughout the entire CL process, dynamic data removal (DDR) to remove less informative training data, and dynamic gradient masking (DGM) to sparsify the gradient updates. Each of them not only improves efficiency, but also further mitigates catastrophic forgetting. SparCL consistently improves the training efficiency of existing state-of-the-art (SOTA) CL methods by at most 23X less training FLOPs, and, surprisingly, further improves the SOTA accuracy by at most 1.7%. SparCL also outperforms competitive baselines obtained from adapting SOTA sparse training methods to the CL setting in both efficiency and accuracy. We also evaluate the effectiveness of SparCL on a real mobile phone, further indicating the practical potential of our method. Source code will be released.

## Action-modulated midbrain dopamine activity arises from distributed control policies

- Jack Lindsey · Ashok Litwin-Kumar
- abstract@[open-review](#): Animal behavior is driven by multiple brain regions working in parallel with distinct control policies. We present a biologically plausible model of off-policy reinforcement learning in the basal ganglia, which enables learning in such an architecture. The model accounts for action-related modulation of dopamine activity that is not captured by previous models that implement on-policy algorithms. In particular, the model predicts that dopamine activity signals a combination of reward prediction error (as in classic models) and "action surprise," a measure of how unexpected an action is relative to the basal ganglia's current policy. In the presence of the action surprise term, the model implements an approximate form of \$Q\$-learning. On benchmark navigation and reaching tasks, we show empirically that this model is capable of learning from data driven completely or in part by other policies (e.g. from other brain regions). By contrast, models without the action surprise term suffer in the presence of additional policies, and are incapable of learning at all from behavior that is completely externally driven. The model provides a computational account for numerous experimental findings about dopamine activity that cannot be explained by classic models of reinforcement learning in the basal ganglia. These include differing levels of action surprise signals in dorsal and ventral striatum, decreasing amounts movement-modulated dopamine activity with practice, and representations of action initiation and kinematics in dopamine activity. It also provides further predictions that can be tested with recordings of striatal dopamine activity.

## The Pitfalls of Regularization in Off-Policy TD Learning

- Gaurav Manek · J. Zico Kolter
- abstract@[open-review](#): Temporal Difference (TD) learning is ubiquitous in reinforcement learning, where it is often combined with off-policy sampling and function approximation. Unfortunately learning with this combination (known as the deadly triad), exhibits instability and unbounded error. To account for this, modern Reinforcement Learning methods often implicitly (or sometimes explicitly) assume that regularization is sufficient to mitigate the problem in practice; indeed, the standard deadly triad examples from the literature can be ``fixed'' via proper regularization. In this paper, we introduce a series of new counterexamples to show that the instability and unbounded error of TD methods is not solved by regularization. We demonstrate that, in the off-policy setting with linear function approximation, TD methods can fail to learn a non-trivial value function under any amount of regularization; we further show that regularization can induce divergence under common conditions; and we show that one of the most promising methods to mitigate this divergence (Emphatic TD algorithms) may also diverge under regularization. We further demonstrate such divergence when using neural networks as function approximators. Thus, we argue that the role of regularization in TD methods needs to be reconsidered, given that it is insufficient to prevent divergence and may itself introduce instability. There needs to be much more care in the practical and theoretical application of regularization to Reinforcement Learning methods.

## VTC-LFC: Vision Transformer Compression with Low-Frequency Components

- Zhenyu Wang · Hao Luo · Pichao WANG · Feng Ding · Fan Wang · Hao Li
- abstract@[open-review](#): Although Vision transformers (ViTs) have recently dominated many vision tasks, deploying ViT models on resource-limited devices remains a challenging problem. To address such a challenge, several methods have been proposed to compress ViTs. Most of them borrow experience in convolutional neural networks (CNNs) and mainly focus on the spatial domain. However, the compression only in the spatial domain suffers from a dramatic performance drop without fine-tuning and is not robust to noise, as the noise in the spatial domain can easily confuse the pruning criteria, leading to some parameters/channels being pruned incorrectly. Inspired by recent findings that self-attention is a low-pass filter and low-frequency signals/components are more informative to ViTs, this paper proposes compressing ViTs with low-frequency components. Two metrics named low-frequency sensitivity (LFS) and low-frequency energy (LFE) are proposed for better channel pruning and token pruning. Additionally, a bottom-up cascade pruning scheme is applied to compress different dimensions jointly. Extensive experiments demonstrate that the proposed method could save 40% ~ 60% of the FLOPs in ViTs, thus significantly increasing the throughput on practical devices with less than 1% performance drop on ImageNet-1K.

## Zeroth-Order Hard-Thresholding: Gradient Error vs. Expansivity

- William de Vazelhes · Hualin Zhang · Huimin Wu · Xiaotong Yuan · Bin Gu
- abstract@[open-review](#): \$\ell\_0\$ constrained optimization is prevalent in machine learning, particularly for high-dimensional problems, because it is a fundamental approach to achieve sparse learning. Hard-thresholding gradient descent is a dominant technique to solve this problem. However, first-order gradients of the objective function may be either unavailable or expensive to calculate in a lot of real-world problems, where zeroth-order (ZO) gradients could be a good surrogate. Unfortunately, whether ZO gradients can work with the hard-thresholding operator is still an unsolved problem. To solve this puzzle, in this paper, we focus on the \$\ell\_0\$ constrained black-box stochastic optimization problems, and propose a new stochastic zeroth-order gradient hard-thresholding (SZOHT) algorithm with a general ZO gradient estimator powered by a novel random support sampling. We provide the convergence analysis of SZOHT under standard assumptions. Importantly, we reveal a conflict between the deviation of ZO estimators and the expansivity of the hard-thresholding operator, and provide a theoretical minimal value of the number of random directions in ZO gradients. In addition, we find that the query complexity of SZOHT is independent or weakly dependent on the dimensionality under different settings. Finally, we illustrate the utility of our method on a portfolio optimization problem as well as black-box adversarial attacks.

## Deep Generative Model for Periodic Graphs

- Shiyu Wang · Xiaojie Guo · Liang Zhao
- abstract@[open-review](#): Periodic graphs are graphs consisting of repetitive local structures, such as crystal nets and polygon mesh. Their generative modeling has great potential in real-world applications such as material design and graphics synthesis. Classical models either rely on domain-specific predefined generation principles (e.g., in crystal net design), or follow geometry-based prescribed rules. Recently, deep generative models have shown great promise in automatically generating general graphs. However, their advancement into periodic graphs has not been well explored due to several key challenges in 1) maintaining graph periodicity; 2) disentangling local and global patterns; and 3) efficiency in learning repetitive patterns. To address them, this paper proposes Periodical-Graph Disentangled Variational Auto-encoder (PGD-VAE), a new deep generative model for periodic graphs that can automatically learn, disentangle, and generate local and global graph patterns. Specifically, we develop a new periodic graph encoder consisting of global-

pattern encoder and local-pattern encoder that ensures to disentangle the representation into global and local semantics. We then propose a new periodic graph decoder consisting of local structure decoder, neighborhood decoder, and global structure decoder, as well as the assembler of their outputs that guarantees periodicity. Moreover, we design a new model learning objective that helps ensure the invariance of local-semantic representations for the graphs with the same local structure. Comprehensive experimental evaluations have been conducted to demonstrate the effectiveness of the proposed method.

## [Finding Second-Order Stationary Points in Nonconvex-Strongly-Concave Minimax Optimization](#)

- Luo Luo · Yujun Li · Cheng Chen
- abstract@[open-review](#): We study the smooth minimax optimization problem  $\min_{\mathbf{x}} \max_{\mathbf{y}} f(\mathbf{x}, \mathbf{y})$ , where  $f$  is  $\ell$ -smooth, strongly-concave in  $\mathbf{y}$  but possibly nonconvex in  $\mathbf{x}$ . Most of existing works focus on finding the first-order stationary point of the function  $f(\mathbf{x}, \mathbf{y})$  or its primal function  $P(\mathbf{x}) \triangleq \max_{\mathbf{y}} f(\mathbf{x}, \mathbf{y})$ , but few of them focus on achieving the second-order stationary point, which is essential to nonconvex problems. In this paper, we propose a novel approach for minimax optimization, called Minimax Cubic Newton (MCN), which could find an  $\mathcal{O}(\kappa^{1.5}\sqrt{\rho}\varepsilon^{-1.5})$ -second-order stationary point of  $P(\mathbf{x})$  with calling  $\mathcal{O}(\kappa^{1.5}\sqrt{\rho}\varepsilon^{-1.5})$  times of second-order oracles and  $\tilde{\mathcal{O}}(\kappa^2\sqrt{\rho}\varepsilon^{-1.5})$  times of first-order oracles, where  $\kappa$  is the condition number and  $\rho$  is the Lipschitz continuous constant for the Hessian of  $f(\mathbf{x}, \mathbf{y})$ . In addition, we propose an inexact variant of MCN for high-dimensional problems to avoid calling the expensive second-order oracles. Instead, our method solves the cubic sub-problem inexactly via gradient descent and matrix Chebyshev expansion. This strategy still obtains the desired approximate second-order stationary point with high probability but only requires  $\tilde{\mathcal{O}}(\kappa^{1.5}\sqrt{\rho}\varepsilon^{-2})$  Hessian-vector oracle calls and  $\tilde{\mathcal{O}}(\kappa^2\sqrt{\rho}\varepsilon^{-1.5})$  first-order oracle calls. To the best of our knowledge, this is the first work that considers the non-asymptotic convergence behavior of finding second-order stationary points for minimax problems without the convex-concave assumptions.

## [Recruitment Strategies That Take a Chance](#)

- Gregory Kehne · Ariel Procaccia · Jingyan Wang
- abstract@[open-review](#): In academic recruitment settings, including faculty hiring and PhD admissions, committees aim to maximize the overall quality of recruited candidates, but there is uncertainty about whether a candidate would accept an offer if given one. Previous work has considered algorithms that make offers sequentially and are subject to a hard budget constraint. We argue that these modeling choices may be inconsistent with the practice of academic recruitment. Instead, we restrict ourselves to a single batch of offers, and we treat the target number of positions as a soft constraint, so we risk overshooting or undershooting the target. Specifically, our objective is to select a subset of candidates that maximizes the overall expected value associated with candidates who accept, minus an expected penalty for deviating from the target. We first analyze the guarantees provided by natural greedy heuristics, showing their desirable properties despite the simplicity. Depending on the structure of the penalty function, we further develop algorithms that provide fully polynomial-time approximation schemes and constant-factor approximations to this objective. Empirical evaluation of our algorithms corroborates these theoretical results.

## [Active Learning of Classifiers with Label and Seed Queries](#)

- Marco Bressan · Nicolò Cesa-Bianchi · Silvio Lattanzi · Andrea Paudice · Maximilian Thiessen
- abstract@[open-review](#): We study exact active learning of binary and multiclass classifiers with margin. Given an  $n$ -point set  $X \subset \mathbb{R}^m$ , we want to learn any unknown classifier on  $X$  whose classes have finite strong convex hull margin, a new notion extending the SVM margin. In the standard active learning setting, where only label queries are allowed, learning a classifier with strong convex hull margin  $\gamma$  requires in the worst case  $\Omega(1/\gamma)^{m-1}$  queries. On the other hand, using the more powerful seed queries (a variant of equivalence queries), the target classifier could be learned in  $O(m \log n)$  queries via Littlestone's Halving algorithm; however, Halving is computationally inefficient. In this work we show that, by carefully combining the two types of queries, a binary classifier can be learned in time  $\operatorname{poly}(n+m)$  using only  $O(m^2 \log n)$  label queries and  $O(m \log \frac{m}{\gamma})$  seed queries; the result extends to  $k$ -class classifiers at the price of a  $k!k^2$  multiplicative overhead. Similar results hold when the input points have bounded bit complexity, or when only one class has strong convex hull margin against the rest. We complement these upper bounds by showing that in the worst case any algorithm needs  $\Omega(\frac{1}{\gamma} \log \frac{1}{\gamma} \log m)$  seed and label queries to learn a  $k$ -class classifier with strong convex hull margin  $\gamma$ .

## [Structure-Aware Image Segmentation with Homotopy Warping](#)

- Xiaoling Hu
- abstract@[open-review](#): Besides per-pixel accuracy, topological correctness is also crucial for the segmentation of images with fine-scale structures, e.g., satellite images and biomedical images. In this paper, by leveraging the theory of digital topology, we identify locations in an image that are critical for topology. By focusing on these critical locations, we propose a new homotopy warping loss to train deep image segmentation networks for better topological accuracy. To efficiently identify these topologically critical locations, we propose a new algorithm exploiting the distance transform. The proposed algorithm, as well as the loss function, naturally generalize to different topological structures in both 2D and 3D settings. The proposed loss function helps deep nets achieve better performance in terms of topology-aware metrics, outperforming state-of-the-art structure/topology-aware segmentation methods.

## [Split-kl and PAC-Bayes-split-kl Inequalities](#)

- Yi-Shan Wu · Yevgeny Seldin
- abstract@[open-review](#): We present a new concentration of measure inequality for sums of independent bounded random variables, which we name a split-kl inequality. The inequality combines the combinatorial power of the kl inequality with ability to exploit low variance. While for Bernoulli random variables the kl inequality is tighter than the Empirical Bernstein, for random variables taking values inside a bounded interval and having low variance the Empirical Bernstein inequality is tighter than the kl. The proposed split-kl inequality yields the best of both worlds. We discuss an application of the split-kl inequality to bounding excess losses. We also derive a PAC-Bayes-split-kl inequality and use a synthetic example and several UCI datasets to compare it with the PAC-Bayes-kl, PAC-Bayes Empirical Bernstein, PAC-Bayes Unexpected Bernstein, and PAC-Bayes Empirical Bennett inequalities.

## [What Makes a "Good" Data Augmentation in Knowledge Distillation - A Statistical Perspective](#)

- Huan Wang · Suhas Lohit · Michael Jones · Yun Fu
- abstract@[open-review](#): Knowledge distillation (KD) is a general neural network training approach that uses a teacher to guide a student. Existing works mainly study KD from the network output side (e.g., how to design a better KD loss function), while few have attempted to understand it from the input side. Especially, its interplay with data augmentation (DA) has not been well understood. In this paper, we ask: Why do some DA schemes (e.g., CutMix) inherently perform much better than others in KD? What characterizes a good DA in KD? Our investigation from a statistical perspective suggests that a good DA scheme should reduce the variance of the teacher's mean probability, which will eventually lead to a lower generalization error gap for the student. Besides the theoretical understanding, we also propose a new entropy-based data-mixing DA scheme to enhance CutMix. Extensive empirical studies support our claims and demonstrate how we can harvest considerable performance gains simply by using a better DA scheme in distillation.

## [Global Optimal K-Medoids Clustering of One Million Samples](#)

- Jiayang Ren · Kaixun Hua · Yankai Cao
- abstract@[open-review](#): We study the deterministic global optimization of the K-Medoids clustering problem. This work proposes a branch and bound (BB) scheme, in which a tailored Lagrangian relaxation method proposed in the 1970s is used to provide a lower bound at each BB node. The lower bounding method already guarantees the maximum gap at the root node. The closed-form solution to the lower bound can be derived analytically without explicitly solving any optimization problems, and its computation can be easily parallelized. Moreover, with this lower bounding method, finite convergence to the global optimal solution can be guaranteed by branching only on the regions of medoids. We also present several tailored bound tightening techniques to reduce the search space and computational cost significantly. Extensive computational studies on 28 machine learning datasets demonstrate that our algorithm can provide a provable global optimal solution with an optimality gap of 0.1% within 4 hours on datasets with up to one million samples. A theoretical proof of global convergence for our algorithm is also presented.

## [SemiFL: Semi-Supervised Federated Learning for Unlabeled Clients with Alternate Training](#)

- Enmao Diao · Jie Ding · Vahid Tarokh
- abstract@[open-review](#): Federated Learning allows the training of machine learning models by using the computation and private data resources of many distributed clients. Most existing results on Federated Learning (FL) assume the clients have ground-truth labels. However, in many practical scenarios, clients may be unable to label task-specific data due to a lack of expertise or resource. We propose SemiFL to address the problem of combining communication efficient FL like FedAvg with Semi-Supervised Learning (SSL). In SemiFL, clients have completely unlabeled data and can train multiple local epochs to reduce communication costs, while the server has a small amount of labeled data. We provide a theoretical understanding of the success of data augmentation-based SSL methods to illustrate the bottleneck of a vanilla combination of communication efficient FL with SSL. To address this issue, we propose alternate training to 'fine-tune global model with labeled data' and 'generate pseudo-labels with global model.' We conduct extensive experiments and demonstrate that our approach significantly improves the performance of a labeled server with unlabeled clients training with multiple local epochs. Moreover, we show that our method outperforms many existing SSFL baselines and performs competitively with the state-of-the-art FL and SSL results.

## [On Optimal Learning Under Targeted Data Poisoning](#)

- Idan Mehalel · Steve Hanneke · Shay Moran · Mohammad Mahmoody · Amin Karbasi
- abstract@[open-review](#): Consider the task of learning a hypothesis class  $\mathcal{H}$  in the presence of an adversary that can replace up to an  $\eta$  fraction of the examples in the training set with arbitrary adversarial examples. The adversary aims to fail the learner on a particular target test point  $x$  which is known to the adversary but not to the learner. In this work we aim to characterize the smallest achievable error  $\epsilon = \epsilon(\eta)$  by the learner in the presence of such an adversary in both realizable and agnostic settings. We fully achieve this in the realizable setting, proving that  $\epsilon = \Theta(\text{VC}(\mathcal{H}) \cdot \eta)$ , where  $\text{VC}(\mathcal{H})$  is the VC dimension of  $\mathcal{H}$ . Remarkably, we show that the upper bound can be attained by a deterministic learner. In the agnostic setting we reveal a more elaborate landscape: we devise a deterministic learner with a multiplicative regret guarantee of  $\epsilon \leq C \cdot \text{OPT} + O(\text{VC}(\mathcal{H}) \cdot \eta)$ , where  $C > 1$  is a universal numerical constant. We complement this by showing that for any deterministic learner there is an attack which worsens its error to at least  $2 \cdot \text{OPT}$ . This implies that a multiplicative deterioration in the regret is unavoidable in this case. Finally, the algorithms we develop for achieving the optimal rates are inherently improper. Nevertheless, we show that for a variety of natural concept classes, such as linear classifiers, it is possible to retain the dependence  $\epsilon = \Theta_{\mathcal{H}}(\eta)$  by a proper algorithm in the realizable setting. Here  $\Theta_{\mathcal{H}}$  conceals a polynomial dependence on  $\text{VC}(\mathcal{H})$ .

## [Disentangling Transfer in Continual Reinforcement Learning](#)

- Maciej Wolczyk · MichaÅ, ZajÄ...c · Razvan Pascanu · Łukasz KuciÅ,ski · Piotr MiÅ,ko
- abstract@[open-review](#): The ability of continual learning systems to transfer knowledge from previously seen tasks in order to maximize forward transfer is a significant challenge for the field, limiting the applicability of continual learning solutions to realistic scenarios. Consequently, this study aims to broaden our understanding of transfer and its driving forces in the specific case of continual reinforcement learning. We adopt SAC as the underlying RL algorithm and Continual World as a suite of continuous control tasks. We systematically study how different components of SAC (the actor and the critic, exploration, and data) affect transfer efficacy, and we provide recommendations regarding various modeling options. The best set of choices, dubbed ClonEx-SAC, is evaluated on the recent Continual World benchmark. ClonEx-SAC achieves 87% final success rate compared to 80% of PackNet, the best method in the benchmark. Moreover, the transfer grows from 0.18 to 0.54 in the metric provided by Continual World.

## [Weak-shot Semantic Segmentation via Dual Similarity Transfer](#)

- Junjie Chen · Li Niu · Siyuan Zhou · Jianlou Si · Chen Qian · Liqing Zhang
- abstract@[open-review](#): Semantic segmentation is a practical and active task, but severely suffers from the expensive cost of pixel-level labels when extending to more classes in wider applications. To this end, we focus on the problem named weak-shot semantic segmentation, where the novel classes are learnt from cheaper image-level labels with the support of base classes having off-the-shelf pixel-level labels. To tackle this problem, we propose a dual similarity transfer framework, which is built upon MaskFormer to disentangle the semantic segmentation task into single-label classification and binary segmentation for each proposal. Specifically, the binary segmentation sub-task allows proposal-pixel similarity transfer from base classes to novel classes, which enables the mask learning of novel classes. We also learn pixel-pixel similarity from base classes and distill such class-agnostic semantic similarity to the semantic masks of novel classes, which regularizes the segmentation model with pixel-level semantic relationship across images. In addition, we propose a complementary loss to facilitate the learning of novel classes. Comprehensive experiments on the challenging COCO-Stuff-10K and ADE20K datasets demonstrate the effectiveness of our method.

## [Intermediate Prototype Mining Transformer for Few-Shot Semantic Segmentation](#)

- YUANWEI LIU · Nian Liu · Xiwen Yao · Junwei Han
- abstract@[open-review](#): Few-shot semantic segmentation aims to segment the target objects in query under the condition of a few annotated support images. Most previous works strive to mine more effective category information from the support to match with the corresponding objects in query. However, they all ignored the category information gap between query and support images. If the objects in them show large intra-class diversity, forcibly migrating the category information from the support to the query is ineffective. To solve this problem, we are the first to introduce an intermediate prototype for mining both deterministic category information from the support and adaptive category knowledge from the query. Specifically, we design an Intermediate Prototype Mining Transformer (IPMT) to learn the prototype in an iterative way. In each IPMT layer, we propagate the object information in both support and query features to the prototype and then use it to activate the query feature map. By conducting this process iteratively, both the intermediate prototype and the query feature can be progressively improved. At last, the final query feature is used to yield precise segmentation prediction. Extensive experiments on both PASCAL-5i and COCO-20i datasets clearly verify the effectiveness of our IPMT and show that it outperforms previous state-of-the-art methods by a large margin. Our code will be released.

## [Towards Safe Reinforcement Learning with a Safety Editor Policy](#)

- Haonan Yu · Wei Xu · Haichao Zhang
- abstract@[open-review](#): We consider the safe reinforcement learning (RL) problem of maximizing utility with extremely low constraint violation rates. Assuming no prior knowledge or pre-training of the environment safety model given a task, an agent has to learn, via exploration, which states and actions are safe. A popular approach in this line of research is to combine a model-free RL algorithm with the Lagrangian method to adjust the weight of the constraint reward relative to the utility reward dynamically. It relies on a single policy to handle the conflict between utility and constraint rewards, which is often challenging. We present SEditor, a two-policy approach that learns a safety editor policy transforming potentially unsafe actions proposed by a utility maximizer policy into safe ones. The safety editor is trained to maximize the constraint reward while minimizing a hinge loss of the utility state-action values before and after an action is edited. SEditor extends existing safety layer designs that assume simplified safety models, to general safe RL scenarios where the safety model can in theory be arbitrarily complex. As a first-order method, it is easy to implement and efficient for both inference and training. On 12 Safety Gym tasks and 2 safe racing tasks, SEditor obtains much a higher overall dominance score than the baselines, and demonstrates outstanding utility performance with constraint violation rates as low as once per 2k time steps, even in obstacle-dense environments. On some tasks, this low violation rate is up to 200 times lower than that of an unconstrained RL method with similar utility performance. Code will be made public.

## [Accelerating SGD for Highly Ill-Conditioned Huge-Scale Online Matrix Completion](#)

- Jialun Zhang · Hong-Ming Chiu · Richard Y Zhang
- abstract@[open-review](#): The matrix completion problem seeks to recover a  $d \times d$  ground truth matrix of low rank  $\|d\|$  from observations of its individual elements. Real-world matrix completion is often a huge-scale optimization problem, with  $d$  so large that even the simplest full-dimension vector operations with  $O(d)$  time complexity become prohibitively expensive. Stochastic gradient descent (SGD) is one of the few algorithms capable of solving matrix completion on a huge scale, and can also naturally handle streaming data over an evolving ground truth. Unfortunately, SGD experiences a dramatic slow-down when the underlying ground truth is ill-conditioned; it requires at least  $O(\kappa \log(1/\epsilon))$  iterations to get  $\epsilon$ -close to ground truth matrix with condition number  $\kappa$ . In this paper, we propose a preconditioned version of SGD that preserves all the favorable practical qualities of SGD for huge-scale online optimization while also making it agnostic to  $\kappa$ . For a symmetric ground truth and the Root Mean Square Error (RMSE) loss, we prove that the preconditioned SGD converges to  $\epsilon$ -accuracy in  $O(\log(1/\epsilon))$  iterations, with a rapid linear convergence rate as if the ground truth were perfectly conditioned with  $\kappa=1$ . In our numerical experiments, we observe a similar acceleration for ill-conditioned matrix completion under the 1-bit cross-entropy loss, as well as pairwise losses such as the Bayesian Personalized Ranking (BPR) loss.

## [Maximum Common Subgraph Guided Graph Retrieval: Late and Early Interaction Networks](#)

- Indradymna Roy · Soumen Chakrabarti · Abir De
- abstract@[open-review](#): The graph retrieval problem is to search in a large corpus of graphs for ones that are most similar to a query graph. A common consideration for scoring similarity is the maximum common subgraph (MCS) between the query and corpus graphs, usually counting the number of common edges (i.e., MCES). In some applications, it is also desirable that the common subgraph be connected, i.e., the maximum common connected subgraph (MCCS). Finding exact MCES and MCCS is intractable, but may be unnecessary if ranking corpus graphs by relevance is the goal. We design fast and trainable neural functions that approximate MCES and MCCS well. Late interaction methods compute dense representations for the query and corpus graph separately, and compare these representations using simple similarity functions at the last stage, leading to highly scalable systems. Early interaction methods combine information from both graphs right from the input stages, are usually considerably more accurate, but slower. We propose both late and early interaction neural MCES and MCCS formulations. They are both based on a continuous relaxation of a node alignment matrix between query and corpus nodes. For MCCS, we propose a novel differentiable network for estimating the size of the largest connected common subgraph. Extensive experiments with seven data sets show that our proposals are superior among late interaction models in terms of both accuracy and speed. Our early interaction models provide accuracy competitive with the state of the art, at substantially greater speeds.

## [Boosting the Performance of Generic Deep Neural Network Frameworks with Log-supermodular CRFs](#)

- Hao Xiong · Yangxiao Lu · Nicholas Ruozzi
- abstract@[open-review](#): Historically, conditional random fields (CRFs) were popular tools in a variety of application areas from computer vision to natural language processing, but due to their higher computational cost and weaker practical performance, they have, in many situations, fallen out of favor and been replaced by end-to-end deep neural network (DNN) solutions. More recently, combined DNN-CRF approaches have been considered, but their speed and practical performance still falls short of the best performing pure DNN solutions. In this work, we present a generic combined approach in which a log-supermodular CRF acts as a regularizer to encourage similarity between outputs in a structured prediction task. We show that this combined approach is widely applicable, practical (it incurs only a moderate overhead on top of the base DNN solution) and, in some cases, it can rival carefully engineered pure DNN solutions for the same structured prediction task.

## [RNNs of RNNs: Recursive Construction of Stable Assemblies of Recurrent Neural Networks](#)

- Leo Kozachkov · Michaela Ennis · Jean-Jacques Slotine
- abstract@[open-review](#): Recurrent neural networks (RNNs) are widely used throughout neuroscience as models of local neural activity. Many properties of single RNNs are well characterized theoretically, but experimental neuroscience has moved in the direction of studying multiple interacting areas, and RNN theory needs to be likewise extended. We take a constructive approach towards this problem, leveraging tools from nonlinear control theory and machine learning to characterize when combinations of stable RNNs will themselves be stable. Importantly, we derive conditions which allow for massive feedback connections between interacting RNNs. We parameterize these conditions for easy optimization using gradient-based techniques, and show that stability-constrained 'network of networks' can perform well on challenging sequential-processing benchmark tasks. Altogether, our results provide a principled approach towards understanding distributed, modular function in the brain.

## [Error Analysis of Tensor-Train Cross Approximation](#)

- Zhen Qin · Alexander Lidiak · Zhixuan Gong · Gongguo Tang · Michael B Wakin · Zhihui Zhu
- abstract@[open-review](#): Tensor train decomposition is widely used in machine learning and quantum physics due to its concise representation of high-dimensional tensors, overcoming the curse of dimensionality. Cross approximation---originally developed for representing a matrix from a set of selected rows and columns---is an efficient method for constructing a tensor train decomposition of a tensor from few of its entries. While tensor train cross approximation has achieved remarkable performance in practical applications, its theoretical analysis, in particular regarding the error of the approximation, is so far lacking. To our knowledge, existing results only provide element-wise approximation accuracy guarantees, which lead to a very loose bound when extended to the entire tensor. In this paper, we bridge this gap by providing accuracy guarantees in terms of the entire tensor for both exact and noisy measurements. Our results illustrate how the choice of selected subtensors affects the quality of the cross approximation and that the approximation error caused by model error and/or measurement error may not grow exponentially with the order of the tensor. These results are verified by numerical experiments, and may have important implications for the usefulness of cross approximations for high-order tensors, such as those encountered in the description of quantum many-body states.

## [Model-based RL with Optimistic Posterior Sampling: Structural Conditions and Sample Complexity](#)

- Alekh Agarwal · Tong Zhang

- abstract@[open-review](#): We propose a general framework to design posterior sampling methods for model-based RL. We show that the proposed algorithms can be analyzed by reducing regret to Hellinger distance based conditional probability estimation. We further show that optimistic posterior sampling can control this Hellinger distance, when we measure model error via data likelihood. This technique allows us to design and analyze unified posterior sampling algorithms with state-of-the-art sample complexity guarantees for many model-based RL settings. We illustrate our general result in many special cases, demonstrating the versatility of our framework.

## Learning interacting dynamical systems with latent Gaussian process ODEs

- [Abstract](#): We study uncertainty-aware modeling of continuous-time dynamics of interacting objects. We introduce a new model that decomposes independent dynamics of single objects accurately from their interactions. By employing latent Gaussian process ordinary differential equations, our model infers both independent dynamics and their interactions with reliable uncertainty estimates. In our formulation, each object is represented as a graph node and interactions are modeled by accumulating the messages coming from neighboring objects. We show that efficient inference of such a complex network of variables is possible with modern variational sparse Gaussian process inference techniques. We empirically demonstrate that our model improves the reliability of long-term predictions over neural network based alternatives and it successfully handles missing dynamic or static information. Furthermore, we observe that only our model can successfully encapsulate independent dynamics and interaction information in distinct functions and show the benefit from this disentanglement in extrapolation scenarios.

## Graph Neural Networks with Adaptive Readouts

- David Buterez · Jon Paul Janet · Steven J Kiddle · Dino Oglic · Pietro Lī<sup>2</sup>
  - abstract@[open-review](#): An effective aggregation of node features into a graph-level representation via readout functions is an essential step in numerous learning tasks involving graph neural networks. Typically, readouts are simple and non-adaptive functions designed such that the resulting hypothesis space is permutation invariant. Prior work on deep sets indicates that such readouts might require complex node embeddings that can be difficult to learn via standard neighborhood aggregation schemes. Motivated by this, we investigate the potential of adaptive readouts given by neural networks that do not necessarily give rise to permutation invariant hypothesis spaces. We argue that in some problems such as binding affinity prediction where molecules are typically presented in a canonical form it might be possible to relax the constraints on permutation invariance of the hypothesis space and learn a more effective model of the affinity by employing an adaptive readout function. Our empirical results demonstrate the effectiveness of neural readouts on more than 40 datasets spanning different domains and graph characteristics. Moreover, we observe a consistent improvement over standard readouts (i.e., sum, max, and mean) relative to the number of neighborhood aggregation iterations and different convolutional operators.

## **Structured Energy Network As a Loss**

- Jay Yoon Lee · Dhruvesh Patel · Purujit Goyal · Wenlong Zhao · Zhiyang Xu · Andrew McCallum
  - abstract@[open-review](#): Belanger & McCallum (2016) and Gygli et al. (2017) have shown that an energy network can capture arbitrary dependencies amongst the output variables in structured prediction; however, their reliance on gradient-based inference (GBI) makes the inference slow and unstable. In this work, we propose Structured Energy As Loss (SEAL) to take advantage of the expressivity of energy networks without incurring the high inference cost. This is a novel learning framework that uses an energy network as a trainable loss function (loss-net) to train a separate neural network (task-net), which is then used to perform the inference through a forward pass. We establish SEAL as a general framework wherein various learning strategies like margin-based, regression, and noise-contrastive, could be employed to learn the parameters of loss-net. Through extensive evaluation on multi-label classification, semantic role labeling, and imagesegmentation, we demonstrate that SEAL provides various useful design choices, is faster at inference than GBI, and leads to significant performance gains over the baselines.

## Data-Driven Conditional Robust Optimization

- Abhilash Reddy Chenreddy · Nymisha Bandi · Erick Delage
  - abstract@[open-review](#): In this paper, we study a novel approach for data-driven decision-making under uncertainty in the presence of contextual information. Specifically, we solve this problem from a Conditional Robust Optimization (CRO) point of view. We propose an integrated framework that designs the conditional uncertainty set by jointly learning the partitions in the covariate data space and simultaneously constructing partition specific deep uncertainty sets for the random vector that perturbs the CRO problem. We also provide theoretical guarantees for the coverage of the uncertainty sets and value at risk performances obtained using the proposed CRO approach. Finally, we use the simulated and real world data to show the implementation of our approach and compare it against two non-contextual benchmark approaches to demonstrate the value of exploiting contextual information in robust optimization.

# Efficient Architecture Search for Diverse Tasks

- Junhong Shen · Misha Khodak · Ameet Talwalkar
  - abstract@[open-review](#): While neural architecture search (NAS) has enabled automated machine learning (AutoML) for well-researched areas, its application to tasks beyond computer vision is still under-explored. As less-studied domains are precisely those where we expect AutoML to have the greatest impact, in this work we study NAS for efficiently solving diverse problems. Seeking an approach that is fast, simple, and broadly applicable, we fix a standard convolutional network (CNN) topology and propose to search for the right kernel sizes and dilations its operations should take on. This dramatically expands the model's capacity to extract features at multiple resolutions for different types of data while only requiring search over the operation space. To overcome the efficiency challenges of naive weight-sharing in this search space, we introduce DASH, a differentiable NAS algorithm that computes the mixture-of-operations using the Fourier diagonalization of convolution, achieving both a better asymptotic complexity and an up-to-10x search time speedup in practice. We evaluate DASH on ten tasks spanning a variety of application domains such as PDE solving, protein folding, and heart disease detection. DASH outperforms state-of-the-art AutoML methods in aggregate, attaining the best-known automated performance on seven tasks. Meanwhile, on six of the ten tasks, the combined search and retraining time is less than 2x slower than simply training a CNN backbone that is far less accurate.

Causal Identification under Markov equivalence: Calculus, Algorithm, and Completeness

- Amin Jaber Å Adele Ribeiro Å Jiji Zhang Å Elias Bareinboim
  - abstract@[open-review](#): One common task in many data sciences applications is to answer questions about the effect of new interventions, like: `what would happen to  $\$Y\$$  if we make  $\$X\$$  equal to  $\$x\$$  while observing covariates  $\$Z=z\$?$ . Formally, this is known as \textit{conditional effect identification}, where the goal is to determine whether a post-interventional distribution is computable from the combination of an observational distribution and assumptions about the underlying domain represented by a causal diagram. A plethora of methods was developed for solving this problem, including the celebrated do-calculus \cite{pearl1995causal}. In practice, these results are not always applicable since they require a fully specified causal diagram as input, which is usually not available. In this paper, we assume as the input of the task a less informative structure known as a partial ancestral graph (PAG), which represents a Markov equivalence class of causal diagrams, learnable from observational data. We make the following contributions under this relaxed setting. First, we introduce a new causal calculus, which subsumes the current state-of-the-art, PAG-calculus. Second, we develop an algorithm for conditional effect identification given a PAG and prove it to be both sound and complete. In words, failure of the algorithm to identify a certain effect implies that this effect is not identifiable by any method. Third, we prove the proposed calculus to be complete for the same task.

## [BILCO: An Efficient Algorithm for Joint Alignment of Time Series](#)

- Xuelong Mi · Mengfan Wang · Alex Chen · Jing-Xuan Lim · Yizhi Wang · Misha B Ahrens · Guoqiang Yu
- abstract@[open-review](#): Multiple time series data occur in many real applications and the alignment among them is usually a fundamental step of data analysis. Frequently, these multiple time series are inter-dependent, which provides extra information for the alignment task and this information cannot be well utilized in the conventional pairwise alignment methods. Recently, the joint alignment was modeled as a max-flow problem, in which both the profile similarity between the aligned time series and the distance between adjacent warping functions are jointly optimized. However, despite the new model having elegant mathematical formulation and superior alignment accuracy, the long computation time and large memory usage, due to the use of the existing general-purpose max-flow algorithms, limit significantly its well-deserved wide use. In this report, we present BIdirectional pushing with Linear Component Operations (BILCO), a novel algorithm that solves the joint alignment max-flow problems efficiently and exactly. We develop the strategy of linear component operations that integrates dynamic programming technique and the push-relabel approach. This strategy is motivated by the fact that the joint alignment max-flow problem is a generalization of dynamic time warping (DTW) and numerous individual DTW problems are embedded. Further, a bidirectional-pushing strategy is proposed to introduce prior knowledge and reduce unnecessary computation, by leveraging another fact that good initialization can be easily computed for the joint alignment max-flow problem. We demonstrate the efficiency of BILCO using both synthetic and real experiments. Tested on thousands of datasets under various simulated scenarios and in three distinct application categories, BILCO consistently achieves at least 10 and averagely 20-folds increase in speed, and uses at most 1/8 and averagely 1/10 memory compared with the best existing max-flow method.

## [Single Model Uncertainty Estimation via Stochastic Data Centering](#)

- Jayaraman Thiagarajan · Rushil Anirudh · Vivek Sivaraman Narayanaswamy · Timo Bremer
- abstract@[open-review](#): We are interested in estimating the uncertainties of deep neural networks, which play an important role in many scientific and engineering problems. In this paper, we present a striking new finding that an ensemble of neural networks with the same weight initialization, trained on datasets that are shifted by a constant bias gives rise to slightly inconsistent trained models, where the differences in predictions are a strong indicator of epistemic uncertainties. Using the neural tangent kernel (NTK), we demonstrate that this phenomena occurs in part because the NTK is not shift-invariant. Since this is achieved via a trivial input transformation, we show that it can therefore be approximated using just a single neural network -- using a technique that we call  $\$Delta\$UQ$  -- that estimates uncertainty around prediction by marginalizing out the effect of the biases. We show that  $\$Delta\$UQ$ 's uncertainty estimates are superior to many of the current methods on a variety of benchmarks-- outlier rejection, calibration under distribution shift, and sequential design optimization of black box functions.

## [Models Out of Line: A Fourier Lens on Distribution Shift Robustness](#)

- Sara Fridovich-Keil · Brian Bartoldson · James Diffenderfer · Bhavya Kailkhura · Timo Bremer
- abstract@[open-review](#): Improving the accuracy of deep neural networks on out-of-distribution (OOD) data is critical to an acceptance of deep learning in real world applications. It has been observed that accuracies on in-distribution (ID) versus OOD data follow a linear trend and models that outperform this baseline are exceptionally rare (and referred to as ``effectively robust''). Recently, some promising approaches have been developed to improve OOD robustness: model pruning, data augmentation, and ensembling or zero-shot evaluating large pretrained models. However, there still is no clear understanding of the conditions on OOD data and model properties that are required to observe effective robustness. We approach this issue by conducting a comprehensive empirical study of diverse approaches that are known to impact OOD robustness on a broad range of natural and synthetic distribution shifts of CIFAR-10 and ImageNet. In particular, we view the "effective robustness puzzle" through a Fourier lens and ask how spectral properties of both models and OOD data correlate with OOD robustness. We find this Fourier lens offers some insight into why certain robust models, particularly those from the CLIP family, achieve OOD robustness. However, our analysis also makes clear that no known metric is consistently the best explanation of OOD robustness. Thus, to aid future research into the OOD puzzle, we address the gap in publicly-available models with effective robustness by introducing a set of pretrained CIFAR-10 models---RobustNets---with varying levels of OOD robustness.

## [Global Convergence and Stability of Stochastic Gradient Descent](#)

- Vivak Patel · Shushu Zhang · Bowen Tian
- abstract@[open-review](#): In machine learning, stochastic gradient descent (SGD) is widely deployed to train models using highly non-convex objectives with equally complex noise models. Unfortunately, SGD theory often makes restrictive assumptions that fail to capture the non-convexity of real problems, and almost entirely ignore the complex noise models that exist in practice. In this work, we demonstrate the restrictiveness of these assumptions using three canonical models in machine learning, then we develop novel theoretical tools to address this shortcoming in two ways. First, we establish that SGD's iterates will either globally converge to a stationary point or diverge under nearly arbitrary nonconvexity and noise models. Under a slightly more restrictive assumption on the joint behavior of the non-convexity and noise model that generalizes current assumptions in the literature, we show that the objective function cannot diverge, even if the iterates diverge. As a consequence of our results, SGD can be applied to a greater range of stochastic optimization problems with confidence about its global convergence behavior and stability.

## [Training Spiking Neural Networks with Local Tandem Learning](#)

- Qu Yang · Jibin Wu · Malu Zhang · Yansong Chua · Xinchao Wang · Haizhou Li
- abstract@[open-review](#): Spiking neural networks (SNNs) have demonstrated great biologically plausibility and energy efficiency over their predecessors. However, there is a lack of an efficient and generalized training method for deep SNNs, especially for deployment on analog computing substrates. In this paper, we put forward a generalized learning rule, termed Local Tandem Learning (LTL). The LTL rule follows the teacher-student learning approach by mimicking the intermediate feature representations of a pre-trained ANN. By decoupling the learning of network layers and leveraging highly informative supervisor signals, we demonstrate rapid network convergence within five training epochs on the CIFAR-10 dataset while having low computational complexity. Our experimental results have also shown that the SNNs thus trained can achieve comparable accuracies to their teacher ANNs on CIFAR-10, CIFAR-100, and Tiny ImageNet datasets. Moreover, the proposed LTL rule is hardware friendly. It can be easily implemented on-chip to perform fast parameter calibration and provide robustness against the notorious device non-ideality issues, including device mismatch, quantization noise, thermal noise, and neuron silencing. It, therefore, opens up a myriad of opportunities for training and deployment of SNN on ultra-low-power mixed-signal neuromorphic computing chips.

## [Robust Calibration with Multi-domain Temperature Scaling](#)

- Yaodong Yu · Stephen Bates · Yi Ma · Michael Jordan
- abstract@[open-review](#): Uncertainty quantification is essential for the reliable deployment of machine learning models to high-stakes application domains. Uncertainty quantification is all the more challenging when training distribution and test distribution are different, even the distribution shifts are mild. Despite the ubiquity of distribution shifts in real-world applications, existing uncertainty quantification approaches mainly study the in-distribution setting where the train and test distributions are the same. In this paper, we develop a systematic calibration model to handle distribution shifts by leveraging data from multiple domains. Our proposed method---multi-domain temperature scaling---uses the heterogeneity in the domains to improve calibration robustness under distribution shift. Through experiments on three benchmark data sets, we find our proposed method outperforms existing methods as measured on both in-distribution and out-of-distribution test sets.

## Roadblocks for Temporarily Disabling Shortcuts and Learning New Knowledge

- Hongjing Niu · Hanting Li · Feng Zhao · Bin Li
- abstract@[open-review](#): Deep learning models have been found with a tendency of relying on shortcuts, i.e., decision rules that perform well on standard benchmarks but fail when transferred to more challenging testing conditions. Such reliance may hinder deep learning models from learning other task-related features and seriously affect their performance and robustness. Although recent studies have shown some characteristics of shortcuts, there are few investigations on how to help the deep learning models to solve shortcut problems. This paper proposes a framework to address this issue by setting up roadblocks on shortcuts. Specifically, it can automatically make modifications based on the original task. Roadblocks are placed during the modification process to ensure that the learned knowledge, including shortcuts, is insufficient to complete the modified task. Therefore, the model trained on the modified task will no longer over-rely on shortcuts. Extensive experiments demonstrate that the proposed framework significantly improves the training of networks on both synthetic and real-world datasets in terms of classification accuracy and feature diversity. Moreover, visualization results show that the mechanism by which our method works is consistent with our expectations. In summary, our approach can effectively disable the shortcuts and thus learn more robust features.

## In What Ways Are Deep Neural Networks Invariant and How Should We Measure This?

- Henry Kvinge · Tegan Emerson · Grayson Jorgenson · Scott Vasquez · Tim Doster · Jesse Lew
- abstract@[open-review](#): It is often said that a deep learning model is ``invariant'' to some specific type of transformation. However, what is meant by this statement strongly depends on the context in which it is made. In this paper we explore the nature of invariance and equivariance of deep learning models with the goal of better understanding the ways that they actually capture these concepts on a formal level. We introduce a family of invariance and equivariance metrics that allow us to quantify these properties in a way that disentangles them from other metrics such as loss or accuracy. We use our metrics to better understand the two most popular methods used to build invariance into networks, data augmentation and equivariant layers. We draw a range of conclusions about invariance and equivariance in deep learning models, ranging from whether initializing a model with pretrained weights has an effect on a trained model's invariance, to the extent to which invariance learned via training can generalize to out-of-distribution data.

## Trust Region Policy Optimization with Optimal Transport Discrepancies: Duality and Algorithm for Continuous Actions

- Antonio Terpin · Nicolas Lanzetti · Batuhan Yardim · Giorgia Ramponi · Florian Dorfler
- abstract@[open-review](#): Policy Optimization (PO) algorithms have been proven particularly suited to handle the high-dimensionality of real-world continuous control tasks. In this context, Trust Region Policy Optimization methods represent a popular approach to stabilize the policy updates. These usually rely on the Kullback-Leibler (KL) divergence to limit the change in the policy. The Wasserstein distance represents a natural alternative, in place of the KL divergence, to define trust regions or to regularize the objective function. However, state-of-the-art works either resort to its approximations or do not provide an algorithm for continuous state-action spaces, reducing the applicability of the method. In this paper, we explore optimal transport discrepancies (which include the Wasserstein distance) to define trust regions, and we propose a novel algorithm - Optimal Transport Trust Region Policy Optimization (OT-TRPO) - for continuous state-action spaces. We circumvent the computational complexity of the infinite-dimensional optimization problem for PO by providing a one-dimensional dual reformulation for which strong duality holds. We then analytically derive the optimal policy update given the solution of the dual problem. This way, we bypass the computation of optimal transport costs and of optimal transport maps, which we implicitly characterize by solving the dual formulation. Finally, we provide an experimental evaluation of our approach across various control tasks. Our results show that optimal transport discrepancies can offer an advantage over state-of-the-art approaches.

## 3DILG: Irregular Latent Grids for 3D Generative Modeling

- Biao Zhang · Matthias Niessner · Peter Wonka
- abstract@[open-review](#): We propose a new representation for encoding 3D shapes as neural fields. The representation is designed to be compatible with the transformer architecture and to benefit both shape reconstruction and shape generation. Existing works on neural fields are grid-based representations with latents being defined on a regular grid. In contrast, we define latents on irregular grids which facilitates our representation to be sparse and adaptive. In the context of shape reconstruction from point clouds, our shape representation built on irregular grids improves upon grid-based methods in terms of reconstruction accuracy. For shape generation, our representation promotes high-quality shape generation using auto-regressive probabilistic models. We show different applications that improve over the current state of the art. First, we show results of probabilistic shape reconstruction from a single higher resolution image. Second, we train a probabilistic model conditioned on very low resolution images. Third, we apply our model to category-conditioned generation. All probabilistic experiments confirm that we are able to generate detailed and high quality shapes to yield the new state of the art in generative 3D shape modeling.

## Dynamic Fair Division with Partial Information

- Gerdus Benade · Daniel Halpern · Alexandros Psomas
- abstract@[open-review](#): We consider the fundamental problem of fairly and efficiently allocating \$T\$ indivisible items among \$n\$ agents with additive preferences. The items become available over a sequence of rounds, and every item must be allocated immediately and irrevocably before the next one arrives. Previous work shows that when the agents' valuations for the items are drawn from known distributions, it is possible (under mild technical assumptions) to find allocations that are envy-free with high probability and Pareto efficient ex-post. We study a \textit{partial-information} setting, where it is possible to elicit ordinal but not cardinal information. When a new item arrives, the algorithm can query each agent for the relative rank of this item with respect to a subset of the past items. When values are drawn from i.i.d.\ distributions, we give an algorithm that is envy-free and \$(1-\epsilon)\$-welfare-maximizing with high probability. We provide similar guarantees (envy-freeness and a constant approximation to welfare with high probability) even with minimally expressive queries that ask for a comparison to a single previous item. For independent but non-identical agents, we obtain envy-freeness and a constant approximation to Pareto efficiency with high probability. We prove that all our results are asymptotically tight.

## Deep Combinatorial Aggregation

- Yuesong Shen · Daniel Cremers
- abstract@[open-review](#): Neural networks are known to produce poor uncertainty estimations, and a variety of approaches have been proposed to remedy this issue. This includes deep ensemble, a simple and effective method that achieves state-of-the-art results for uncertainty-aware learning tasks. In this work, we explore a combinatorial generalization of deep ensemble called deep combinatorial aggregation (DCA). DCA creates multiple instances of network components and aggregates their combinations to produce diversified model proposals and predictions. DCA components can be defined at different levels of granularity. And we discovered that coarse-grain DCAs can outperform deep ensemble for uncertainty-aware learning both in terms of predictive performance and uncertainty estimation. For fine-grain DCAs, we discover that an average parameterization approach named deep combinatorial weight averaging (DCWA) can improve the baseline training. It is on par with stochastic weight averaging (SWA) but does not require any custom training schedule or adaptation of BatchNorm layers. Furthermore, we propose a consistency enforcing loss that helps the training of DCWA and modelwise DCA. We experiment on in-domain, distributional shift, and out-of-distribution image classification tasks, and empirically confirm the effectiveness of DCWA and DCA approaches.

## Fairness Reprogramming

- Guanhua Zhang · Yihua Zhang · Yang Zhang · Wenqi Fan · Qing Li · Sijia Liu · Shiyu Chang
- abstract@[open-review](#): Despite a surge of recent advances in promoting machine Learning (ML) fairness, the existing mainstream approaches mostly require training or finetuning the entire weights of the neural network to meet the fairness criteria. However, this is often infeasible in practice for those large-scale trained models due to large computational and storage costs, low data efficiency, and model privacy issues. In this paper, we propose a new generic fairness learning paradigm, called FairReprogram, which incorporates the model reprogramming technique. Specifically, FairReprogram considers the neural model fixed, and instead appends to the input a set of perturbations, called the fairness trigger, which is tuned towards the fairness criteria under a min-max formulation. We further introduce an information-theoretic framework that explains why and under what conditions fairness goals can be achieved using the fairness trigger. We show both theoretically and empirically that the fairness trigger can effectively obscure demographic biases in the output prediction of fixed ML models by providing false demographic information that hinders the model from utilizing the correct demographic information to make the prediction. Extensive experiments on both NLP and CV datasets demonstrate that our method can achieve better fairness improvements than retraining-based methods with far less training cost and data dependency under two widely-used fairness criteria.

## [Subspace clustering in high-dimensions: Phase transitions \& Statistical-to-Computational gap](#)

- Luca Pesce · Bruno Loureiro · Florent Krzakala · Lenka Zdeborová;
- abstract@[open-review](#): A simple model to study subspace clustering is the high-dimensional  $k$ -Gaussian mixture model where the cluster means are sparse vectors. Here we provide an exact asymptotic characterization of the statistically optimal reconstruction error in this model in the high-dimensional regime with extensive sparsity, i.e. when the fraction of non-zero components of the cluster means  $\rho$ , as well as the ratio  $\alpha$  between the number of samples and the dimension are fixed, while the dimension diverges. We identify the information-theoretic threshold below which obtaining a positive correlation with the true cluster means is statistically impossible. Additionally, we investigate the performance of the approximate message passing (AMP) algorithm analyzed via its state evolution, which is conjectured to be optimal among polynomial algorithm for this task. We identify in particular the existence of a statistical-to-computational gap between the algorithm that requires a signal-to-noise ratio  $\lambda_{\text{alg}} \geq k / \sqrt{\alpha}$  to perform better than random, and the information theoretic threshold at  $\lambda_{\text{it}} \approx \sqrt{-k \rho \log \rho} / \sqrt{\alpha}$ . Finally, we discuss the case of sub-extensive sparsity  $\rho$  by comparing the performance of the AMP with other sparsity-enhancing algorithms, such as sparse-PCA and diagonal thresholding.

## [Sound and Complete Incorporation of Local Causal Background Knowledge with Latent Variables](#)

- Tian-Zuo Wang · Tian Qin · Zhi-Hua Zhou
- abstract@[open-review](#): Identifying causal relations is an important problem in various disciplines of science. When latent variables exist, ancestral graph is generally used to describe causal relations. Only a Markov equivalence class (MEC) is identifiable with observational data, represented by a partial ancestral graph (PAG) where the orientations of some edges are uncertain. Without further assumptions, we can only eliminate the uncertainty by incorporating background knowledge attainable through experiments or human experience. However, it is an open problem how to completely orient a PAG with background knowledge. The problem is fundamental due to its implication for the maximally identifiable causal relations based on the information. In this paper, we take a further step towards addressing it. We notice that the background knowledge in real tasks is usually in a local form. Given local background knowledge, we present sound and complete orientation rules, with which the PAG can be maximally oriented. As an application of the orientation rules, we propose the first general active learning framework for causal discovery in the presence of latent confounders, aiming to recover the true ancestral graph with as few interventions as possible. Specifically, we present a baseline maximal entropy criterion, equipped with Metropolis-Hastings sampling, to select the interventional target in each round. The experiments demonstrate the effectiveness and efficiency of the proposed framework to discover the ancestral graph.

## [Sampling in Constrained Domains with Orthogonal-Space Variational Gradient Descent](#)

- Ruqi Zhang · Qiang Liu · Xin Tong
- abstract@[open-review](#): Sampling methods, as important inference and learning techniques, are typically designed for unconstrained domains. However, constraints are ubiquitous in machine learning problems, such as those on safety, fairness, robustness, and many other properties that must be satisfied to apply sampling results in real-life applications. Enforcing these constraints often leads to implicitly-defined manifolds, making efficient sampling with constraints very challenging. In this paper, we propose a new variational framework with a designed orthogonal-space gradient flow (O-Gradient) for sampling on a manifold  $\mathcal{G}_0$  defined by general equality constraints. O-Gradient decomposes the gradient into two parts: one decreases the distance to  $\mathcal{G}_0$  and the other decreases the KL divergence in the orthogonal space. While most existing manifold sampling methods require initialization on  $\mathcal{G}_0$ , O-Gradient does not require such prior knowledge. We prove that O-Gradient converges to the target constrained distribution with rate  $\tilde{O}(1/\text{number of iterations})$  under mild conditions. Our proof relies on a new Stein characterization of conditional measure which could be of independent interest. We implement O-Gradient through both Langevin dynamics and Stein variational gradient descent and demonstrate its effectiveness in various experiments, including Bayesian deep neural networks.

## [Generalization Bounds for Stochastic Gradient Descent via Localized \$\bar{\epsilon}\$ -Covers](#)

- Sejun Park · Umut Simsekli · Murat Erdogdu
- abstract@[open-review](#): In this paper, we propose a new covering technique localized for the trajectories of SGD. This localization provides an algorithm-specific complexity measured by the covering number, which can have dimension-independent cardinality in contrast to standard uniform covering arguments that result in exponential dimension dependency. Based on this localized construction, we show that if the objective function is a finite perturbation of a piecewise strongly convex and smooth function with  $P$  pieces, i.e., non-convex and non-smooth in general, the generalization error can be upper bounded by  $O(\sqrt{(\log n \log(nP))/n})$ , where  $n$  is the number of data samples. In particular, this rate is independent of dimension and does not require early stopping and decaying step size. Finally, we employ these results in various contexts and derive generalization bounds for multi-index linear models, multi-class support vector machines, and  $K$ -means clustering for both hard and soft label setups, improving the previously known state-of-the-art rates.

## [Multi-Fidelity Best-Arm Identification](#)

- Riccardo Poiani · Alberto Maria Metelli · Marcello Restelli
- abstract@[open-review](#): In several real-world applications, a learner has access to multiple environment simulators, each with a different precision (e.g., simulation accuracy) and cost (e.g., computational time). In such a scenario, the learner faces the trade-off between selecting expensive accurate simulators or preferring cheap imprecise ones. We formalize this setting as a multi-fidelity variant of the stochastic best-arm identification problem, where querying the original arm is expensive, but multiple and biased approximations (i.e., fidelities) are available at lower costs. The learner's goal, in this setting, is to sequentially choose which simulator to query in order to minimize the total cost, while guaranteeing to identify the optimal arm with high probability. We first derive a lower bound on the identification cost, assuming that the maximum bias of each fidelity is known to the learner. Then, we propose a novel algorithm, Iterative Imprecise Successive Elimination (IISE), which provably reduces the total cost w.r.t. algorithms that ignore the multi-fidelity structure and whose cost complexity upper bound mimics the structure of the lower bound. Furthermore, we show that the cost complexity of IISE can be further reduced when the agent has access to a more fine-grained knowledge of the error introduced by the approximators. Finally, we numerically validate IISE, showing the benefits of our method in simulated domains.

## [An efficient graph generative model for navigating ultra-large combinatorial synthesis libraries](#)

- Aryan Pedawi Å· Paweł Gniewek Å· Chaoyi Chang Å· Brandon Anderson Å· Henry van den Bedem
- abstract@[open-review](#): Virtual, make-on-demand chemical libraries have transformed early-stage drug discovery by unlocking vast, synthetically accessible regions of chemical space. Recent years have witnessed rapid growth in these libraries from millions to trillions of compounds, hiding undiscovered, potent hits for a variety of therapeutic targets. However, they are quickly approaching a size beyond that which permits explicit enumeration, presenting new challenges for virtual screening. To overcome these challenges, we propose the Combinatorial Synthesis Library Variational Auto-Encoder (CSLVAE). The proposed generative model represents such libraries as a differentiable, hierarchically-organized database. Given a compound from the library, the molecular encoder constructs a query for retrieval, which is utilized by the molecular decoder to reconstruct the compound by first decoding its chemical reaction and subsequently decoding its reactants. Our design minimizes autoregression in the decoder, facilitating the generation of large, valid molecular graphs. Our method performs fast and parallel batch inference for ultra-large synthesis libraries, enabling a number of important applications in early-stage drug discovery. Compounds proposed by our method are guaranteed to be in the library, and thus synthetically and cost-effectively accessible. Importantly, CSLVAE can encode out-of-library compounds and search for in-library analogues. In experiments, we demonstrate the capabilities of the proposed method in the navigation of massive combinatorial synthesis libraries.

## [Learning to Generate Inversion-Resistant Model Explanations](#)

- Hoyong Jeong Å· Suyoung Lee Å· Sung Ju Hwang Å· Sooel Son
- abstract@[open-review](#): The wide adoption of deep neural networks (DNNs) in mission-critical applications has spurred the need for interpretable models that provide explanations of the model's decisions. Unfortunately, previous studies have demonstrated that model explanations facilitate information leakage, rendering DNN models vulnerable to model inversion attacks. These attacks enable the adversary to reconstruct original images based on model explanations, thus leaking privacy-sensitive features. To this end, we present Generative Noise Injector for Model Explanations (GNIME), a novel defense framework that perturbs model explanations to minimize the risk of model inversion attacks while preserving the interpretabilities of the generated explanations. Specifically, we formulate the defense training as a two-player minimax game between the inversion attack network on the one hand, which aims to invert model explanations, and the noise generator network on the other, which aims to inject perturbations to tamper with model inversion attacks. We demonstrate that GNIME significantly decreases the information leakage in model explanations, decreasing transferable classification accuracy in facial recognition models by up to 84.8% while preserving the original functionality of model explanations.

## [Towards a holistic assessment of health data representations under realistic dataset shifts](#)

- Neeraj Wagh Å· Jionghao Wei Å· Samarth Rawal Å· Brent M Berry Å· Yogatheesan Varatharajah
- abstract@[open-review](#): The recent availability of large datasets in bio-medicine has inspired the development of representation learning methods for multiple healthcare applications. Despite advances in predictive performance, the clinical utility of such methods is limited when exposed to real-world data. Here we develop model diagnostic measures to detect potential pitfalls during deployment without assuming access to external data. Specifically, we focus on modeling realistic data shifts in electrophysiological signals (EEGs) via data transforms, and extend the conventional task-based evaluations with analyses of a) model's latent space and b) predictive uncertainty, under these transforms. We conduct experiments on multiple EEG feature encoders and two clinically relevant downstream tasks using publicly available large-scale clinical EEGs. Within this experimental setting, our results suggest that measures of latent space integrity and model uncertainty under the proposed data shifts may help anticipate performance degradation during deployment.

## [Not too little, not too much: a theoretical analysis of graph \(over\)smoothing](#)

- Nicolas Keriven
- abstract@[open-review](#): We analyze graph smoothing with mean aggregation, where each node successively receives the average of the features of its neighbors. Indeed, it has quickly been observed that Graph Neural Networks (GNNs), which generally follow some variant of Message-Passing (MP) with repeated aggregation, may be subject to the oversmoothing phenomenon: by performing too many rounds of MP, the node features tend to converge to a non-informative limit. In the case of mean aggregation, for connected graphs, the node features become constant across the whole graph. At the other end of the spectrum, it is intuitively obvious that some MP rounds are necessary, but existing analyses do not exhibit both phenomena at once: beneficial ``finite'' smoothing and oversmoothing in the limit. In this paper, we consider simplified linear GNNs, and rigorously analyze two examples for which a finite number of mean aggregation steps provably improves the learning performance, before oversmoothing kicks in. We consider a latent space random graph model, where node features are partial observations of the latent variables and the graph contains pairwise relationships between them. We show that graph smoothing restores some of the lost information, up to a certain point, by two phenomenon: graph smoothing shrinks non-principal directions in the data faster than principal ones, which is useful for regression, and shrinks nodes within communities faster than they collapse together, which improves classification.

## [DivBO: Diversity-aware CASH for Ensemble Learning](#)

- Yu Shen Å· Yupeng Lu Å· Yang Li Å· Yaofeng Tu Å· Wentao Zhang Å· Bin CUI
- abstract@[open-review](#): The Combined Algorithm Selection and Hyperparameters optimization (CASH) problem is one of the fundamental problems in Automated Machine Learning (AutoML). Motivated by the success of ensemble learning, recent AutoML systems build post-hoc ensembles to output the final predictions instead of using the best single learner. However, while most CASH methods focus on searching for a single learner with the best performance, they neglect the diversity among base learners (i.e., they may suggest similar configurations to previously evaluated ones), which is also a crucial consideration when building an ensemble. To tackle this issue and further enhance the ensemble performance, we propose DivBO, a diversity-aware framework to inject explicit search of diversity into the CASH problems. In the framework, we propose to use a diversity surrogate to predict the pair-wise diversity of two unseen configurations. Furthermore, we introduce a temporary pool and a weighted acquisition function to guide the search of both performance and diversity based on Bayesian optimization. Empirical results on 15 public datasets show that DivBO achieves the best average ranks (1.82 and 1.73) on both validation and test errors among 10 compared methods, including post-hoc designs in recent AutoML systems and state-of-the-art baselines for ensemble learning on CASH problems.

## [Adam Can Converge Without Any Modification On Update Rules](#)

- Yushun Zhang Å· Congliang Chen Å· Naichen Shi Å· Ruoyu Sun Å· Zhi-Quan Luo
- abstract@[open-review](#): Ever since \citet{reddi2019convergence} pointed out the divergence issue of Adam, many new variants have been designed to obtain convergence. However, vanilla Adam remains exceptionally popular and it works well in practice. Why is there a gap between theory and practice? We point out there is a mismatch between the settings of theory and practice: \citet{reddi2019convergence} pick the problem after picking the hyperparameters of Adam, i.e.,  $(\beta_1, \beta_2)$ ; while practical applications often fix the problem first and then tune  $(\beta_1, \beta_2)$ . Due to this observation, we conjecture that the empirical convergence can be theoretically justified, only if we change the order of picking the problem and hyperparameter. In this work, we confirm this conjecture. We prove that, when the 2nd-order momentum parameter  $\beta_2$  is large and 1st-order momentum parameter  $\beta_1 < \sqrt{\beta_2} < 1$ , Adam converges to the neighborhood of critical points. The size of the neighborhood is proportional to the variance of stochastic gradients. Under an extra condition (strong growth condition), Adam converges to critical points. As  $\beta_2$  increases, our convergence result can cover any  $\beta_1 \in [0, 1]$  including  $\beta_1 = 0.9$ , which is the default setting in deep learning libraries. To our knowledge, this is the first result showing that Adam can converge under a wide range of hyperparameters without any modification on its update rules. Further, our analysis does not require assumptions of bounded gradients or bounded 2nd-order momentum. When  $\beta_2$  is small, we further point out a large region of  $(\beta_1, \beta_2)$  combinations where Adam can diverge to infinity. Our divergence result considers the same setting (fixing the optimization problem ahead) as our convergence result, indicating that there is a phase transition from divergence to convergence when increasing  $\beta_2$ . These

positive and negative results provide suggestions on how to tune Adam hyperparameters: for instance, when Adam does not work well, we suggest tuning up  $\beta_2$  and trying  $\beta_1 < \sqrt{\beta_2}$ .

## [Towards Efficient Post-training Quantization of Pre-trained Language Models](#)

- Haoli Bai · Lu Hou · Lifeng Shang · Xin Jiang · Irwin King · Michael R Lyu
- abstract@[open-review](#): Network quantization has gained increasing attention with the rapid growth of large pre-trained language models~(PLMs). However, most existing quantization methods for PLMs follow quantization-aware training~(QAT) that requires end-to-end training with full access to the entire dataset. Therefore, they suffer from slow training, large memory overhead, and data accessibility issues. In this paper, we study post-training quantization~(PTQ) of PLMs, and propose module-wise quantization error minimization~(MREM), an efficient solution to mitigate these issues. By partitioning the PLM into multiple modules, we minimize the reconstruction error incurred by quantization for each module. In addition, we design a new model parallel training strategy such that each module can be trained locally on separate computing devices without waiting for preceding modules, which brings nearly the theoretical training speed-up (e.g.,  $4\times$  on  $4$  GPUs). Experiments on GLUE and SQuAD benchmarks show that our proposed PTQ solution not only performs close to QAT, but also enjoys significant reductions in training time, memory overhead, and data consumption.

## [Iterative Scene Graph Generation](#)

- Siddhesh Khandelwal · Leonid Sigal
- abstract@[open-review](#): The task of scene graph generation entails identifying object entities and their corresponding interaction predicates in a given image (or video). Due to the combinatorially large solution space, existing approaches to scene graph generation assume certain factorization of the joint distribution to make the estimation feasible (e.g., assuming that objects are conditionally independent of predicate predictions). However, this fixed factorization is not ideal under all scenarios (e.g., for images where an object entailed in interaction is small and not discernible on its own). In this work, we propose a novel framework for scene graph generation that addresses this limitation, as well as introduces dynamic conditioning on the image, using message passing in a Markov Random Field. This is implemented as an iterative refinement procedure wherein each modification is conditioned on the graph generated in the previous iteration. This conditioning across refinement steps allows joint reasoning over entities and relations. This framework is realized via a novel and end-to-end trainable transformer-based architecture. In addition, the proposed framework can improve existing approach performance. Through extensive experiments on Visual Genome and Action Genome benchmark datasets we show improved performance on the scene graph generation.

## [Task-level Differentially Private Meta Learning](#)

- Xinyu Zhou · Raef Bassily
- abstract@[open-review](#): We study the problem of meta-learning with task-level differential privacy. Meta-learning has received increasing attention recently because of its ability to enable fast generalization to new task with small number of data points. However, the training process of meta learning likely involves exchange of task specific information, which may pose privacy risk especially in some privacy-sensitive applications. Therefore, it is important to provide strong privacy guarantees such that the learning process will not reveal any task sensitive information. To this end, existing works have proposed meta learning algorithms with record-level differential privacy, which is not sufficient in many scenarios since it does not protect the aggregated statistics based on the task dataset as a whole. Moreover, the utility guarantees in the prior work are based on assuming that the loss function satisfies both smoothness and quadratic growth conditions, which do not necessarily hold in practice. To address these issues, we propose meta learning algorithms with task-level differential privacy; that is, our algorithms protect the privacy of the entire dataset for each task. In the case when a single meta model is trained, we give both privacy and utility guarantees assuming only that the loss is convex and Lipschitz. Moreover, we propose a new private clustering-based meta-learning algorithm that enables private meta learning of multiple meta models. This can provide significant accuracy gains over the single meta model paradigm, especially when the tasks distribution cannot be well represented by a single meta model. Finally, we conduct several experiments demonstrating the effectiveness of our proposed algorithms.

## [Diagonal State Spaces are as Effective as Structured State Spaces](#)

- Ankit Gupta · Albert Gu · Jonathan Berant
- abstract@[open-review](#): Modeling long range dependencies in sequential data is a fundamental step towards attaining human-level performance in many modalities such as text, vision, audio and video. While attention-based models are a popular and effective choice in modeling short-range interactions, their performance on tasks requiring long range reasoning has been largely inadequate. In an exciting result, Gu et al. (ICLR 2022) proposed the `Structured State Space` (S4) architecture delivering large gains over state-of-the-art models on several long-range tasks across various modalities. The core proposition of S4 is the parameterization of state matrices via a diagonal plus low rank structure, allowing efficient computation. In this work, we show that one can match the performance of S4 even without the low rank correction and thus assuming the state matrices to be diagonal. Our `Diagonal State Space` (DSS) model matches the performance of S4 on Long Range Arena tasks, speech classification on Speech Commands dataset, while being conceptually simpler and straightforward to implement.

## [Structural Analysis of Branch-and-Cut and the Learnability of Gomory Mixed Integer Cuts](#)

- Maria-Florina Balcan · Siddharth Prasad · Tuomas Sandholm · Ellen Vitercik
- abstract@[open-review](#): The incorporation of cutting planes within the branch-and-bound algorithm, known as branch-and-cut, forms the backbone of modern integer programming solvers. These solvers are the foremost method for solving discrete optimization problems and thus have a vast array of applications in machine learning, operations research, and many other fields. Choosing cutting planes effectively is a major research topic in the theory and practice of integer programming. We conduct a novel structural analysis of branch-and-cut that pins down how every step of the algorithm is affected by changes in the parameters defining the cutting planes added to the input integer program. Our main application of this analysis is to derive sample complexity guarantees for using machine learning to determine which cutting planes to apply during branch-and-cut. These guarantees apply to infinite families of cutting planes, such as the family of Gomory mixed integer cuts, which are responsible for the main breakthrough speedups of integer programming solvers. We exploit geometric and combinatorial structure of branch-and-cut in our analysis, which provides a key missing piece for the recent generalization theory of branch-and-cut.

## [Bidirectional Learning for Offline Infinite-width Model-based Optimization](#)

- Can Chen · Yingxueff Zhang · Jie Fu · Xue (Steve) Liu · Mark Coates
- abstract@[open-review](#): In offline model-based optimization, we strive to maximize a black-box objective function by only leveraging a static dataset of designs and their scores. This problem setting arises in numerous fields including the design of materials, robots, DNAs, proteins, etc. Recent approaches train a deep neural network (DNN) model on the static dataset to act as a proxy function, and then perform gradient ascent on the existing designs to obtain potentially high-scoring designs. This methodology frequently suffers from the out-of-distribution problem where the proxy function often returns adversarial designs. To mitigate this problem, we propose  $\text{BID}^2$ —bidirectional learning for offline infinite-width model-based optimization.  $\text{BID}^2$  consists of two mappings: the forward mapping leverages the static dataset to predict the scores of the high-scoring designs, and the backward mapping leverages the high-scoring designs to predict the scores of the static dataset. The backward mapping, neglected in previous work, can distill more information of the static dataset into the high-scoring designs, which effectively mitigates the out-of-distribution problem. Yet, for a finite-width DNN model, the loss function of the backward mapping is intractable and only has an approximate form, which leads to a

significant deterioration of the design quality. We thus adopt an infinite-width DNN model and propose to employ the corresponding neural tangent kernel to yield a closed-form loss for more accurate design updates. Experiments on various tasks verify the effectiveness of BDI. The code will be available upon paper acceptance.

## [Path Independent Equilibrium Networks Can Better Exploit Test-Time Computation](#)

- Cem Anil · Ashwini Pokle · Kaiqu Liang · Johannes Treutlein · Yuhuai Wu · Shaojie Bai · J. Zico Kolter · Roger Grosse
- abstract@[open-review](#): Designing networks capable of attaining better performance with an increased inference budget is important to facilitate generalization to harder problem instances. Recent efforts have shown promising results in this direction by making use of depth-wise recurrent networks. In this work, we reproduce the performance of the prior art using a broader class of architectures called equilibrium models, and find that stronger generalization performance on harder examples (which require more iterations of inference to get correct) strongly correlates with the path independence of the system—its ability to converge to the same attractor (or limit cycle) regardless of initialization, given enough computation. Experimental interventions made to promote path independence result in improved generalization on harder (and thus more compute-hungry) problem instances, while those that penalize it degrade this ability. Path independence analyses are also useful on a per-example basis: for equilibrium models that have good in-distribution performance, path independence on out-of-distribution samples strongly correlates with accuracy. Thus, considering equilibrium models and path independence jointly leads to a valuable new viewpoint under which we can study the generalization performance of these networks on hard problem instances.

## [On the convergence of policy gradient methods to Nash equilibria in general stochastic games](#)

- Angeliki Giannou · Kyriakos Lotidis · Panayotis Mertikopoulos · Emmanouil-Vasileios Vlatakis-Gkaragkounis
- abstract@[open-review](#): Multi-agent learning in stochastic  $N$ -player games is a notoriously difficult problem because, in addition to their changing strategic decisions, the players of the game must also contend with the fact that the game itself evolves over time, possibly in a very complicated manner. Because of this, the equilibrium convergence properties of popular learning algorithms — like policy gradient and its variants — are poorly understood, except in specific classes of games (such as potential or two-player, zero-sum games). In view of all this, we examine the long-run behavior of policy gradient methods with respect to Nash equilibrium policies that are second-order stationary (SOS) in a sense similar to the type of KKT sufficiency conditions used in optimization. Our analysis shows that SOS policies are locally attracting with high probability, and we show that policy gradient trajectories with gradient estimates provided by the REINFORCE algorithm achieve an  $\mathcal{O}(1/\sqrt{n})$  convergence rate to such equilibria if the method's step-size is chosen appropriately. On the other hand, when the equilibrium in question is deterministic, we show that this rate can be improved dramatically and, in fact, policy gradient methods converge within a finite number of iterations in that case.

## [Conformal Frequency Estimation with Sketched Data](#)

- Matteo Sesia · Stefano Favaro
- abstract@[open-review](#): A flexible conformal inference method is developed to construct confidence intervals for the frequencies of queried objects in very large data sets, based on a much smaller sketch of those data. The approach is data-adaptive and requires no knowledge of the data distribution or of the details of the sketching algorithm; instead, it constructs provably valid frequentist confidence intervals under the sole assumption of data exchangeability. Although our solution is broadly applicable, this paper focuses on applications involving with the count-min sketch algorithm and a non-linear variation thereof. The performance is compared to that of frequentist and Bayesian alternatives through simulations and experiments with data sets of SARS-CoV-2 DNA sequences and classic English literature.

## [Single Loop Gaussian Homotopy Method for Non-convex Optimization](#)

- Hidenori Iwakiri · Yuhang Wang · Shinji Ito · Akiko Takeda
- abstract@[open-review](#): The Gaussian homotopy (GH) method is a popular approach to finding better stationary points for non-convex optimization problems by gradually reducing a parameter value  $t$ , which changes the problem to be solved from an almost convex one to the original target one. Existing GH-based methods repeatedly call an iterative optimization solver to find a stationary point every time  $t$  is updated, which incurs high computational costs. We propose a novel single loop framework for GH methods (SLGH) that updates the parameter  $t$  and the optimization decision variables at the same. Computational complexity analysis is performed on the SLGH algorithm under various situations: either a gradient or gradient-free oracle of a GH function can be obtained for both deterministic and stochastic settings. The convergence rate of SLGH with a tuned hyperparameter becomes consistent with the convergence rate of gradient descent, even though the problem to be solved is gradually changed due to  $t$ . In numerical experiments, our SLGH algorithms show faster convergence than an existing double loop GH method while outperforming gradient descent-based methods in terms of finding a better solution.

## [Deep Surrogate Assisted Generation of Environments](#)

- Varun Bhatt · Bryon Tjanaka · Matthew Fontaine · Stefanos Nikolaidis
- abstract@[open-review](#): Recent progress in reinforcement learning (RL) has started producing generally capable agents that can solve a distribution of complex environments. These agents are typically tested on fixed, human-authored environments. On the other hand, quality diversity (QD) optimization has been proven to be an effective component of environment generation algorithms, which can generate collections of high-quality environments that are diverse in the resulting agent behaviors. However, these algorithms require potentially expensive simulations of agents on newly generated environments. We propose Deep Surrogate Assisted Generation of Environments (DSAGE), a sample-efficient QD environment generation algorithm that maintains a deep surrogate model for predicting agent behaviors in new environments. Results in two benchmark domains show that DSAGE significantly outperforms existing QD environment generation algorithms in discovering collections of environments that elicit diverse behaviors of a state-of-the-art RL agent and a planning agent.

## [Constraining Gaussian Processes to Systems of Linear Ordinary Differential Equations](#)

- Andreas Besginow · Markus Lange-Hegermann
- abstract@[open-review](#): Data in many applications follows systems of Ordinary Differential Equations (ODEs). This paper presents a novel algorithmic and symbolic construction for covariance functions of Gaussian Processes (GPs) with realizations strictly following a system of linear homogeneous ODEs with constant coefficients, which we call LODE-GPs. Introducing this strong inductive bias into a GP improves modelling of such data. Using smith normal form algorithms, a symbolic technique, we overcome two current restrictions in the state of the art: (1) the need for certain uniqueness conditions in the set of solutions, typically assumed in classical ODE solvers and their probabilistic counterparts, and (2) the restriction to controllable systems, typically assumed when encoding differential equations in covariance functions. We show the effectiveness of LODE-GPs in a number of experiments, for example learning physically interpretable parameters by maximizing the likelihood.

## [Predictive Querying for Autoregressive Neural Sequence Models](#)

- Alex Boyd · Samuel Showalter · Stephan Mandt · Padhraic Smyth

- abstract@[open-review](#): In reasoning about sequential events it is natural to pose probabilistic queries such as "when will event A occur next" or "what is the probability of A occurring before B," with applications in areas such as user modeling, medicine, and finance. However, with machine learning shifting towards neural autoregressive models such as RNNs and transformers, probabilistic querying has been largely restricted to simple cases such as next-event prediction. This is in part due to the fact that future querying involves marginalization over large path spaces, which is not straightforward to do efficiently in such models. In this paper we introduce a general typology for predictive queries in autoregressive sequence models and show that such queries can be systematically represented by sets of elementary building blocks. We leverage this typology to develop new query estimation methods based on beam search, importance sampling, and hybrids. Across four large-scale sequence datasets from different application domains, as well as for the GPT-2 language model, we demonstrate the ability to make query answering tractable for arbitrary queries in exponentially-large predictive path-spaces, and find clear differences in cost-accuracy tradeoffs between search and sampling methods.

## [Joint Model-Policy Optimization of a Lower Bound for Model-Based RL](#)

- Benjamin Eysenbach · Alexander Khazatsky · Sergey Levine · Russ Salakhutdinov
- abstract@[open-review](#): Many model-based reinforcement learning (RL) methods follow a similar template: fit a model to previously observed data, and then use data from that model for RL or planning. However, models that achieve better training performance (e.g., lower MSE) are not necessarily better for control: an RL agent may seek out the small fraction of states where an accurate model makes mistakes, or it might act in ways that do not expose the errors of an inaccurate model. As noted in prior work, there is an objective mismatch: models are useful if they yield good policies, but they are trained to maximize their accuracy, rather than the performance of the policies that result from them. In this work, we propose a single objective for jointly training the model and the policy, such that updates to either component increase a lower bound on expected return. To the best of our knowledge, this is the first lower bound for model-based RL that holds globally and can be efficiently estimated in continuous settings; it is the only lower bound that mends the objective mismatch problem. A version of this bound becomes tight under certain assumptions. Optimizing this bound resembles a GAN: a classifier distinguishes between real and fake transitions, the model is updated to produce transitions that look realistic, and the policy is updated to avoid states where the model predictions are unrealistic. Numerical simulations demonstrate that optimizing this bound yields reward maximizing policies and yields dynamics that (perhaps surprisingly) can aid in exploration. We also show that a deep RL algorithm loosely based on our lower bound can achieve performance competitive with prior model-based methods, and better performance on certain hard exploration tasks.

## [Coresets for Vertical Federated Learning: Regularized Linear Regression and \\$K\\$-Means Clustering](#)

- Lingxiao Huang · Zhize Li · Jialin Sun · Haoyu Zhao
- abstract@[open-review](#): Vertical federated learning (VFL), where data features are stored in multiple parties distributively, is an important area in machine learning. However, the communication complexity for VFL is typically very high. In this paper, we propose a unified framework by constructing \texttt{coresets} in a distributed fashion for communication-efficient VFL. We study two important learning tasks in the VFL setting: regularized linear regression and \$k\$-means clustering, and apply our coreset framework to both problems. We theoretically show that using coresets can drastically alleviate the communication complexity, while nearly maintain the solution quality. Numerical experiments are conducted to corroborate our theoretical findings.

## [Analyzing Sharpness along GD Trajectory: Progressive Sharpening and Edge of Stability](#)

- Zixuan Wang · Zhouzi Li · Jian Li
- abstract@[open-review](#): Recent findings (e.g., \textit{cohen2021gradient}) demonstrate that modern neural networks trained by full-batch gradient descent typically enter a regime called Edge of Stability (EOS). In this regime, the sharpness, i.e., the maximum Hessian eigenvalue, first increases to the value  $2/(\text{step size})$  (the progressive sharpening phase) and then oscillates around this value (the EOS phase). This paper aims to analyze the GD dynamics and the sharpness along the optimization trajectory. Our analysis naturally divides the GD trajectory into four phases depending on the change of the sharpness. We empirically identify the norm of output layer weight as an interesting indicator of sharpness dynamics. Based on this empirical observation, we attempt to theoretically and empirically explain the dynamics of various key quantities that lead to the change of sharpness in each phase of EOS. Moreover, based on certain assumptions, we provide a theoretical proof of the sharpness behavior in EOS regime in two-layer fully-connected linear neural networks. We also discuss some other empirical findings and the limitation of our theoretical results.

## [MoCoDA: Model-based Counterfactual Data Augmentation](#)

- Silviu Pitis · Elliot Creager · Ajay Mandlekar · Animesh Garg
- abstract@[open-review](#): The number of states in a dynamic process is exponential in the number of objects, making reinforcement learning (RL) difficult in complex, multi-object domains. For agents to scale to the real world, they will need to react to and reason about unseen combinations of objects. We argue that the ability to recognize and use local factorization in transition dynamics is a key element in unlocking the power of multi-object reasoning. To this end, we show that (1) known local structure in the environment transitions is sufficient for an exponential reduction in the sample complexity of training a dynamics model, and (2) a locally factored dynamics model provably generalizes out-of-distribution to unseen states and actions. Knowing the local structure also allows us to predict which unseen states and actions this dynamics model will generalize to. We propose to leverage these observations in a novel Model-based Counterfactual Data Augmentation (MoCoDA) framework. MoCoDA applies a learned locally factored dynamics model to an augmented distribution of states and actions to generate counterfactual transitions for RL. MoCoDA works with a broader set of local structures than prior work and allows for direct control over the augmented training distribution. We show that MoCoDA enables RL agents to learn policies that generalize to unseen states and actions. We use MoCoDA to train an offline RL agent to solve an out-of-distribution robotics manipulation task on which standard offline RL algorithms fail.

## [ALMA: Hierarchical Learning for Composite Multi-Agent Tasks](#)

- Shariq Iqbal · Robby Costales · Fei Sha
- abstract@[open-review](#): Despite significant progress on multi-agent reinforcement learning (MARL) in recent years, coordination in complex domains remains a challenge. Work in MARL often focuses on solving tasks where agents interact with all other agents and entities in the environment; however, we observe that real-world tasks are often composed of several isolated instances of local agent interactions (subtasks), and each agent can meaningfully focus on one subtask to the exclusion of all else in the environment. In these composite tasks, successful policies can often be decomposed into two levels of decision-making: agents are allocated to specific subtasks and each agent acts productively towards their assigned subtask alone. This decomposed decision making provides a strong structural inductive bias, significantly reduces agent observation spaces, and encourages subtask-specific policies to be reused and composed during training, as opposed to treating each new composition of subtasks as unique. We introduce ALMA, a general learning method for taking advantage of these structured tasks. ALMA simultaneously learns a high-level subtask allocation policy and low-level agent policies. We demonstrate that ALMA learns sophisticated coordination behavior in a number of challenging environments, outperforming strong baselines. ALMA's modularity also enables it to better generalize to new environment configurations. Finally, we find that while ALMA can integrate separately trained allocation and action policies, the best performance is obtained only by training all components jointly. Our code is available at <https://github.com/shariqiqbal2810/ALMA>

## [Coordinates are not lonely - Codebook Prior Helps Implicit Neural 3D representations](#)

- Fukun Yin · Wen Liu · Zilong Huang · Pei Cheng · Tao Chen · Gang Yu

- abstract@[open-review](#): Implicit neural 3D representation has achieved impressive results in surface or scene reconstruction and novel view synthesis, which typically uses the coordinate-based multi-layer perceptrons (MLPs) to learn a continuous scene representation. However, existing approaches, such as Neural Radiance Field (NeRF) and its variants, usually require dense input views (i.e. 50-150) to obtain decent results. To relieve the over-dependence on massive calibrated images and enrich the coordinate-based feature representation, we explore injecting the prior information into the coordinate-based network and introduce a novel coordinate-based model, CoCo-INR, for implicit neural 3D representation. The cores of our method are two attention modules: codebook attention and coordinate attention. The former extracts the useful prototypes containing rich geometry and appearance information from the prior codebook, and the latter propagates such prior information into each coordinate and enriches its feature representation for a scene or object surface. With the help of the prior information, our method can render 3D views with more photo-realistic appearance and geometries than the current methods using fewer calibrated images available. Experiments on various scene reconstruction datasets, including DTU and BlendedMVS, and the full 3D head reconstruction dataset, H3DS, demonstrate the robustness under fewer input views and fine detail-preserving capability of our proposed method.

## [Scalable and Efficient Training of Large Convolutional Neural Networks with Differential Privacy](#)

- Zhiqi Bu Â· Jialin Mao Â· Shiyun Xu
- abstract@[open-review](#): Large convolutional neural networks (CNN) can be difficult to train in the differentially private (DP) regime, since the optimization algorithms require a computationally expensive operation, known as the per-sample gradient clipping. We propose an efficient and scalable implementation of this clipping on convolutional layers, termed as the mixed ghost clipping, that significantly eases the private training in terms of both time and space complexities, without affecting the accuracy. The improvement in efficiency is rigorously studied through the first complexity analysis for the mixed ghost clipping and existing DP training algorithms. Extensive experiments on vision classification tasks, with large ResNet, VGG, and Vision Transformers (ViT), demonstrate that DP training with mixed ghost clipping adds  $\$1 \sim 10\%$  memory overhead and  $\$<2\times$  slowdown to the standard non-private training. Specifically, when training VGG19 on CIFAR10, the mixed ghost clipping is  $\$3\times$  faster than state-of-the-art Opacus library with  $\$18\times$  larger maximum batch size. To emphasize the significance of efficient DP training on convolutional layers, we achieve 96.7% accuracy on CIFAR10 and 83.0% on CIFAR100 at  $\$epsilon=1$  using BEiT, while the previous best results are 94.8% and 67.4%, respectively. We open-source a privacy engine ([url{https://github.com/woodyx218/private\\_vision}](https://github.com/woodyx218/private_vision)) that implements DP training of CNN (including convolutional ViT) with a few lines of code.

## [Retaining Knowledge for Learning with Dynamic Definition](#)

- Zichang Liu Â· Benjamin Coleman Â· Tianyi Zhang Â· Anshumali Shrivastava
- abstract@[open-review](#): Machine learning models are often deployed in settings where they must be constantly updated in response to the changes in class definitions while retaining high accuracy on previous learned definitions. A classical use case is the task of fraud detection, where what can be considered as fraud keeps evolving over time. While such an update can be accomplished by re-training on the complete data, the process is inefficient and prevents real-time and on-device learning. On the other hand, efficient methods that incrementally learn from new data often result in the forgetting of previously-learned knowledge. We formally define this problem as Learning with Dynamic Definition (LDD) and demonstrate that popular models, such as the Vision Transformer and Roberta, exhibit substantial forgetting of past definitions. We present a first practical and provable solution to LDD. Our proposal is a hash-based memory model `RIDDLE` that solves evolving definitions by associating different instances only to relevant parameters. We prove that our model is a universal function approximator and theoretically bound the knowledge lost during the update process. On practical tasks with evolving class definition in vision and NLP, our model outperforms baselines by up to 30% on the original dataset while providing competitive accuracy on the update dataset.

## [Exploring evolution-based & -free protein language models as protein function predictors](#)

- Mingyang Hu Â· Fajie Yuan Â· Kevin Yang Â· Fusong Ju Â· Jin Su Â· Hui Wang Â· Fei Yang Â· Qiuyang Ding
- abstract@[open-review](#): Large-scale Protein Language Models (PLMs) have improved performance in protein prediction tasks, ranging from 3D structure prediction to various function predictions. In particular, AlphaFold, a ground-breaking AI system, could potentially reshape structural biology. However, the utility of the PLM module in AlphaFold, Evoformer, has not been explored beyond structure prediction. In this paper, we investigate the representation ability of three popular PLMs: ESM-1b (single sequence), MSA-Transformer (multiple sequence alignment), and Evoformer (structural), with a special focus on Evoformer. Specifically, we aim to answer the following key questions: (1) Does the Evoformer trained as part of AlphaFold produce representations amenable to predicting protein function? (2) If yes, can Evoformer replace ESM-1b and MSA-Transformer? (3) How much do these PLMs rely on evolution-related protein data? In this regard, are they complementary to each other? We compare these models by empirical study along with new insights and conclusions. Finally, we release code and datasets for reproducibility.

## [MGNNI: Multiscale Graph Neural Networks with Implicit Layers](#)

- Juncheng Liu Â· Bryan Hooi Â· Kenji Kawaguchi Â· Xiaokui Xiao
- abstract@[open-review](#): Recently, implicit graph neural networks (GNNs) have been proposed to capture long-range dependencies in underlying graphs. In this paper, we introduce and justify two weaknesses of implicit GNNs: the constrained expressiveness due to their limited effective range for capturing long-range dependencies, and their lack of ability to capture multiscale information on graphs at multiple resolutions. To show the limited effective range of previous implicit GNNs, We first provide a theoretical analysis and point out the intrinsic relationship between the effective range and the convergence of iterative equations used in these models. To mitigate the mentioned weaknesses, we propose a multiscale graph neural network with implicit layers (MGNNI) which is able to model multiscale structures on graphs and has an expanded effective range for capturing long-range dependencies. We conduct comprehensive experiments for both node classification and graph classification to show that MGNNI outperforms representative baselines and has a better ability for multiscale modeling and capturing of long-range dependencies.

## [A Robust Phased Elimination Algorithm for Corruption-Tolerant Gaussian Process Bandits](#)

- Ilija Bogunovic Â· Zihan Li Â· Andreas Krause Â· Jonathan Scarlett
- abstract@[open-review](#): We consider the sequential optimization of an unknown, continuous, and expensive to evaluate reward function, from noisy and adversarially corrupted observed rewards. When the corruption attacks are subject to a suitable budget  $\$C\$$  and the function lives in a Reproducing Kernel Hilbert Space (RKHS), the problem can be posed as `{em corrupted Gaussian process (GP) bandit optimization}`. We propose a novel robust elimination-type algorithm that runs in epochs, combines exploration with infrequent switching to select a small subset of actions, and plays each action for multiple time instants. Our algorithm, `{em Robust GP Phased Elimination (RGP-PE)}`, successfully balances robustness to corruptions with exploration and exploitation such that its performance degrades minimally in the presence (or absence) of adversarial corruptions. When  $\$T\$$  is the number of samples and  $\$gamma_T\$$  is the maximal information gain, the corruption-dependent term in our regret bound is  $\$O(C \gamma_T^{3/2})\$$ , which is significantly tighter than the existing  $\$O(C \sqrt{T \gamma_T})\$$  for several commonly-considered kernels. We perform the first empirical study of robustness in the corrupted GP bandit setting, and show that our algorithm is robust against a variety of adversarial attacks.

## [Amortized Mixing Coupling Processes for Clustering](#)

- Huafeng Liu Â· Liping Jing
- abstract@[open-review](#): Considering the ever-increasing scale of data, which may contain tens of thousands of data points or complicated latent structures, the issue of scalability and algorithmic efficiency becomes of vital importance for clustering. In this paper, we propose cluster-wise amortized mixing

coupling processes (AMCP), which is able to achieve efficient amortized clustering in a well-defined non-parametric Bayesian posterior. Specifically, AMCP learns clusters sequentially with the aid of the proposed intra-cluster mixing (IntraCM) and inter-cluster coupling (InterCC) strategies, which investigate the relationship between data points and reference distribution in a linear optimal transport mixing view, and coupling the unassigned set and assigned set to generate new cluster. IntraCM and InterCC avoid pairwise calculation of distances between clusters and reduce the computational complexity from quadratic to linear in the current number of clusters. Furthermore, cluster-wise sequential process is able to improve the quick adaptation ability for the next cluster generation. In this case, AMCP simultaneously learns what makes a cluster, how to group data points into clusters, and how to adaptively control the number of clusters. To illustrate the superiority of the proposed method, we perform experiments on both synthetic data and real-world data in terms of clustering performance and computational efficiency.

## [On the Adversarial Robustness of Mixture of Experts](#)

- Joan Puigcerver · Rodolphe Jenatton · Carlos Riquelme · Pranjal Awasthi · Srinadh Bhojanapalli
- abstract@[open-review](#): Adversarial robustness is a key desirable property of neural networks. It has been empirically shown to be affected by their sizes, with larger networks being typically more robust. Recently, \citet{bubeck2021universal} proved a lower bound on the Lipschitz constant of functions that fit the training data in terms of their number of parameters. This raises an interesting open question, do---and can---functions with more parameters, but not necessarily more computational cost, have better robustness? We study this question for sparse Mixture of Expert models (MoEs), that make it possible to scale up the model size for a roughly constant computational cost. We theoretically show that under certain conditions on the routing and the structure of the data, MoEs can have significantly smaller Lipschitz constants than their dense counterparts. The robustness of MoEs can suffer when the highest weighted experts for an input implement sufficiently different functions. We next empirically evaluate the robustness of MoEs on ImageNet using adversarial attacks and show they are indeed more robust than dense models with the same computational cost. We make key observations showing the robustness of MoEs to the choice of experts, highlighting the redundancy of experts in models trained in practice.

## [First-Order Algorithms for Min-Max Optimization in Geodesic Metric Spaces](#)

- Michael Jordan · Tianyi Lin · Emmanouil-Vasileios Vlatakis-Gkaragkounis
- abstract@[open-review](#): From optimal transport to robust dimensionality reduction, many machine learning applications can be cast into the min-max optimization problems over Riemannian manifolds. Though many min-max algorithms have been analyzed in the Euclidean setting, it has been elusive how these results translate to the Riemannian case. Zhang et al. (2022) have recently identified that geodesic convex-concave Riemannian problems admit always Sionâ€™s saddle point solutions. Immediately, an important question that arises is if a performance gap between the Riemannian and the optimal Euclidean space convex-concave algorithms is necessary. Our work is the first to answer the question in the negative: We prove that the Riemannian corrected extragradient (RCEG) method achieves last-iterate at a linear convergence rate at the geodesically strongly convex-concave case, matching the Euclidean one. Our results also extend to the stochastic or non-smooth case where RCEG & Riemannian gradient-ascent descent (RGDA) achieve respectively near-optimal convergence rates up to factors depending on curvature of the manifold. Finally, we empirically demonstrate the effectiveness of RCEG in solving robust PCA.

## [Toward Efficient Robust Training against Union of \$\ell\_p\$ Threat Models](#)

- Gaurang Sriramanan · Maharsi Gor · Soheil Feizi
- abstract@[open-review](#): The overwhelming vulnerability of deep neural networks to carefully crafted perturbations known as adversarial attacks has led to the development of various training techniques to produce robust models. While the primary focus of existing approaches has been directed toward addressing the worst-case performance achieved under a single-threat model, it is imperative that safety-critical systems are robust with respect to multiple threat models simultaneously. Existing approaches that address worst-case performance under the union of such threat models ( $\ell_\infty$ ,  $\ell_2$ ,  $\ell_1$ ) either utilize adversarial training methods that require multi-step attacks which are computationally expensive in practice, or rely upon fine-tuning of pre-trained models that are robust with respect to a single-threat model. In this work, we show that by carefully choosing the objective function used for robust training, it is possible to achieve similar, or improved worst-case performance over a union of threat models while utilizing only single-step attacks, thereby achieving a significant reduction in computational resources necessary for training. Furthermore, prior work showed that adversarial training specific to the  $\ell_1$  threat model is relatively difficult, to the extent that even multi-step adversarially trained models were shown to be prone to gradient-masking. However, the proposed methodâ€”when applied on the  $\ell_1$  threat model specificallyâ€”enables us to obtain the first  $\ell_1$  robust model trained solely with single-step adversaries. Finally, to demonstrate the merits of our approach, we utilize a modern set of attack evaluations to better estimate the worst-case performance under the considered union of threat models.

## [Trajectory-guided Control Prediction for End-to-end Autonomous Driving: A Simple yet Strong Baseline](#)

- Penghao Wu · Xiaosong Jia · Li Chen · Junchi Yan · Hongyang Li · Yu Qiao
- abstract@[open-review](#): Current end-to-end autonomous driving methods either run a controller based on a planned trajectory or perform control prediction directly, which have spanned two separately studied lines of research. Seeing their potential mutual benefits to each other, this paper takes the initiative to explore the combination of these two well-developed worlds. Specifically, our integrated approach has two branches for trajectory planning and direct control, respectively. The trajectory branch predicts the future trajectory, while the control branch involves a novel multi-step prediction scheme such that the relationship between current actions and future states can be reasoned. The two branches are connected so that the control branch receives corresponding guidance from the trajectory branch at each time step. The outputs from two branches are then fused to achieve complementary advantages. Our results are evaluated in the closed-loop urban driving setting with challenging scenarios using the CARLA simulator. Even with a monocular camera input, the proposed approach ranks first on the official CARLA Leaderboard, outperforming other complex candidates with multiple sensors or fusion mechanisms by a large margin.

## [On the detrimental effect of invariances in the likelihood for variational inference](#)

- Richard Kurle · Ralf Herbrich · Tim Januschowski · Yuyang (Bernie) Wang · Jan Gasthaus
- abstract@[open-review](#): Variational Bayesian posterior inference often requires simplifying approximations such as mean-field parametrisation to ensure tractability. However, prior work has associated the variational mean-field approximation for Bayesian neural networks with underfitting in the case of small datasets or large model sizes. In this work, we show that invariances in the likelihood function of over-parametrised models contribute to this phenomenon because these invariances complicate the structure of the posterior by introducing discrete and/or continuous modes which cannot be well approximated by Gaussian mean-field distributions. In particular, we show that the mean-field approximation has an additional gap in the evidence lower bound compared to a purpose-built posterior that takes into account the known invariances. Importantly, this invariance gap is not constant; it vanishes as the approximation reverts to the prior. We proceed by first considering translation invariances in a linear model with a single data point in detail. We show that, while the true posterior can be constructed from a mean-field parametrisation, this is achieved only if the objective function takes into account the invariance gap. Then, we transfer our analysis of the linear model to neural networks. Our analysis provides a framework for future work to explore solutions to the invariance problem.

## [Truncated proposals for scalable and hassle-free simulation-based inference](#)

- Michael Deistler · Pedro Goncalves · Jakob H Macke

- abstract@[open-review](#): Simulation-based inference (SBI) solves statistical inverse problems by repeatedly running a stochastic simulator and inferring posterior distributions from model-simulations. To improve simulation efficiency, several inference methods take a sequential approach and iteratively adapt the proposal distributions from which model simulations are generated. However, many of these sequential methods are difficult to use in practice, both because the resulting optimisation problems can be challenging and efficient diagnostic tools are lacking. To overcome these issues, we present Truncated Sequential Neural Posterior Estimation (TSNPE). TSNPE performs sequential inference with truncated proposals, sidestepping the optimisation issues of alternative approaches. In addition, TSNPE allows to efficiently perform coverage tests that can scale to complex models with many parameters. We demonstrate that TSNPE performs on par with previous methods on established benchmark tasks. We then apply TSNPE to two challenging problems from neuroscience and show that TSNPE can successfully obtain the posterior distributions, whereas previous methods fail. Overall, our results demonstrate that TSNPE is an efficient, robust, and accurate inference method that can scale to problems that were previously inaccessible to neural posterior estimation.

## [Local-Global MCMC kernels: the best of both worlds](#)

- Sergey Samsonov · Evgeny Lagutin · Marylou Gabriëls · Alain Durmus · Alexey Naumov · Eric Moulines
- abstract@[open-review](#): Recent works leveraging learning to enhance sampling have shown promising results, in particular by designing effective non-local moves and global proposals. However, learning accuracy is inevitably limited in regions where little data is available such as in the tails of distributions as well as in high-dimensional problems. In the present paper we study an Explore-Exploit Markov chain Monte Carlo strategy ( $\text{\$operatorname{Ex}^2MCMC}$ ) that combines local and global samplers showing that it enjoys the advantages of both approaches. We prove  $\$V$ -uniform geometric ergodicity of  $\text{\$operatorname{Ex}^2MCMC}$  without requiring a uniform adaptation of the global sampler to the target distribution. We also compute explicit bounds on the mixing rate of the Explore-Exploit strategy under realistic conditions. Moreover, we propose an adaptive version of the strategy ( $\text{\$operatorname{FIE}^2MCMC}$ ) where a normalizing flow is trained while sampling to serve as a proposal for global moves. We illustrate the efficiency of  $\text{\$operatorname{Ex}^2MCMC}$  and its adaptive version on classical sampling benchmarks as well as in sampling high-dimensional distributions defined by Generative Adversarial Networks seen as Energy Based Models.

## [Spectral Bias in Practice: The Role of Function Frequency in Generalization](#)

- Sara Fridovich-Keil · Raphael Gontijo Lopes · Rebecca Roelofs
- abstract@[open-review](#): Despite their ability to represent highly expressive functions, deep learning models seem to find simple solutions that generalize surprisingly well. Spectral bias -- the tendency of neural networks to prioritize learning low frequency functions -- is one possible explanation for this phenomenon, but so far spectral bias has primarily been observed in theoretical models and simplified experiments. In this work, we propose methodologies for measuring spectral bias in modern image classification networks on CIFAR-10 and ImageNet. We find that these networks indeed exhibit spectral bias, and that interventions that improve test accuracy on CIFAR-10 tend to produce learned functions that have higher frequencies overall but lower frequencies in the vicinity of examples from each class. This trend holds across variation in training time, model architecture, number of training examples, data augmentation, and self-distillation. We also explore the connections between function frequency and image frequency and find that spectral bias is sensitive to the low frequencies prevalent in natural images. On ImageNet, we find that learned function frequency also varies with internal class diversity, with higher frequencies on more diverse classes. Our work enables measuring and ultimately influencing the spectral behavior of neural networks used for image classification, and is a step towards understanding why deep models generalize well.

## [Geometry-aware Two-scale PIFu Representation for Human Reconstruction](#)

- Zheng Dong · Ke Xu · Ziheng Duan · Hujun Bao · Weiwei Xu · Rynson Lau
- abstract@[open-review](#): Although PIFu-based 3D human reconstruction methods are popular, the quality of recovered details is still unsatisfactory. In a sparse (e.g., 3 RGBD sensors) capture setting, the depth noise is typically amplified in the PIFu representation, resulting in flat facial surfaces and geometry-fallible bodies. In this paper, we propose a novel geometry-aware two-scale PIFu for 3D human reconstruction from sparse, noisy inputs. Our key idea is to exploit the complementary properties of depth denoising and 3D reconstruction, for learning a two-scale PIFu representation to reconstruct high-frequency facial details and consistent bodies separately. To this end, we first formulate depth denoising and 3D reconstruction as a multi-task learning problem. The depth denoising process enriches the local geometry information of the reconstruction features, while the reconstruction process enhances depth denoising with global topology information. We then propose to learn the two-scale PIFu representation using two MLPs based on the denoised depth and geometry-aware features. Extensive experiments demonstrate the effectiveness of our approach in reconstructing facial details and bodies of different poses and its superiority over state-of-the-art methods.

## [ZeroQuant: Efficient and Affordable Post-Training Quantization for Large-Scale Transformers](#)

- Zhewei Yao · Reza Yazdani Aminabadi · Minjia Zhang · Xiaoxia Wu · Conglong Li · Yuxiong He
- abstract@[open-review](#): How to efficiently serve ever-larger trained natural language models in practice has become exceptionally challenging even for powerful cloud servers due to their prohibitive memory/computation requirements. In this work, we present an efficient and affordable post-training quantization approach to compress large Transformer-based models, termed as \OURS. \OURS is an end-to-end quantization and inference pipeline with three main components: (1) a fine-grained hardware-friendly quantization scheme for both weight and activations; (2) a novel affordable layer-by-layer knowledge distillation algorithm (\lwd) even without the original training data access;(3) a highly-optimized quantization system backend support to remove the quantization/dequantization overhead. As such, we are able to show that:(1) \OURS can reduce the precision for weight and activations to INT8 in a cost-free way for both \bert and \gpt-style models with minimal accuracy impact, which leads to up to 5.19x/4.16x speedup on \bert\gpt-style models compared to FP16 inference, separately;(2) \OURS plus \lwd can affordably quantize the weights in the fully-connected module to INT4 along with INT8 weights in the attention module and INT8 activations, resulting in 3x memory footprint reduction compared to the FP16 model;(3) \OURS can be directly applied to two of the largest open-sourced language models, including \gptneox, for which our INT8 model achieves similar accuracy as the FP16 model but achieves 5.2x better efficiency. Our code is open-sourced at \cite{code\_compression}.

## [ResQ: A Residual Q Function-based Approach for Multi-Agent Reinforcement Learning Value Factorization](#)

- Siqi SHEN · Mengwei Qiu · Jun Liu · Weiquan Liu · Yongquan Fu · Xinwang Liu · Cheng Wang
- abstract@[open-review](#): The factorization of state-action value functions for Multi-Agent Reinforcement Learning (MARL) is important. Existing studies are limited by their representation capability, sample efficiency, and approximation error. To address these challenges, we propose, ResQ, a MARL value function factorization method, which can find the optimal joint policy for any state-action value function through residual functions. ResQ masks some state-action value pairs from a joint state-action value function, which is transformed as the sum of a main function and a residual function. ResQ can be used with mean-value and stochastic-value RL. We theoretically show that ResQ can satisfy both the individual global max (IGM) and the distributional IGM principle without representation limitations. Through experiments on matrix games, the predator-prey, and StarCraft benchmarks, we show that ResQ can obtain better results than multiple expected/stochastic value factorization methods.

## [Diversified Recommendations for Agents with Adaptive Preferences](#)

- William Brown · Arpit Agarwal
- abstract@[open-review](#): When an Agent visits a platform recommending a menu of content to select from, their choice of item depends not only on immutable preferences, but also on their prior engagements with the platform. The Recommender's primary objective is typically to encourage content

consumption which optimizes some reward, such as ad revenue, but they often additionally aim to ensure that a sufficiently wide variety of content is consumed by the Agent over time. We formalize this problem as an adversarial bandit task. At each step, the Recommender presents a menu of  $\$k\$$  (out of  $\$n\$$ ) items to the Agent, who selects one item in the menu according to their unknown {item preference model}, which maps their history of past items to relative selection probabilities. The Recommender then observes the Agent's selected item and receives bandit feedback of the item's (adversarial) reward. In addition to optimizing reward from the selected items at each step, the Recommender must also ensure that the total distribution of chosen items has sufficiently high entropy. We define a class of preference models which are {item locally learnable}, i.e.\ behavior over the entire domain can be estimated by only observing behavior in a small region; this includes models representable by bounded-degree polynomials as well as functions with a sparse Fourier basis. For this class, we give an algorithm for the Recommender which obtains  $\$Tilde{O}(T^{3/4})\$$  regret against all item distributions satisfying two conditions: they are sufficiently diversified, and they are {item instantaneously realizable} at any history by some distribution over menus. We show that these conditions are closely connected: all sufficiently high-entropy distributions are instantaneously realizable at any history of selected items. We also give a set of negative results justifying our assumptions, in the form of a runtime lower bound for non-local learning and linear regret lower bounds for alternate benchmarks.

## [WT-MVSNet: Window-based Transformers for Multi-view Stereo](#)

- Jinli Liao · Yikang Ding · Yoli Shavit · Dihe Huang · Shihao Ren · Jia Guo · Kai Zhang · Wensen Feng
- abstract@[open-review](#): Recently, Transformers were shown to enhance the performance of multi-view stereo by enabling long-range feature interaction. In this work, we propose Window-based Transformers (WT) for local feature matching and global feature aggregation in multi-view stereo. We introduce a Window-based Epipolar Transformer (WET) which reduces matching redundancy by using epipolar constraints. Since point-to-line matching is sensitive to erroneous camera pose and calibration, we match windows near the epipolar lines. A second Shifted WT is employed for aggregating global information within cost volume. We present a novel Cost Transformer (CT) to replace 3D convolutions for cost volume regularization. In order to better constrain the estimated depth maps from multiple views, we further design a novel geometric consistency loss (Geo Loss) which punishes unreliable areas where multi-view consistency is not satisfied. Our WT multi-view stereo method (WT-MVSNet) achieves state-of-the-art performance across multiple datasets and ranks  $\$1^{st}\$$  on Tanks and Temples benchmark. Code will be available upon acceptance.

## [Understanding Cross-Domain Few-Shot Learning Based on Domain Similarity and Few-Shot Difficulty](#)

- Jaehoon Oh · Sungnyun Kim · Namgyu Ho · Jin-Hwa Kim · Hwanjun Song · Se-Young Yun
- abstract@[open-review](#): Cross-domain few-shot learning (CD-FSL) has drawn increasing attention for handling large differences between the source and target domains--an important concern in real-world scenarios. To overcome these large differences, recent works have considered exploiting small-scale unlabeled data from the target domain during the pre-training stage. This data enables self-supervised pre-training on the target domain, in addition to supervised pre-training on the source domain. In this paper, we empirically investigate which pre-training is preferred based on domain similarity and few-shot difficulty of the target domain. We discover that the performance gain of self-supervised pre-training over supervised pre-training becomes large when the target domain is dissimilar to the source domain, or the target domain itself has low few-shot difficulty. We further design two pre-training schemes, mixed-supervised and two-stage learning, that improve performance. In this light, we present six findings for CD-FSL, which are supported by extensive experiments and analyses on three source and eight target benchmark datasets with varying levels of domain similarity and few-shot difficulty. Our code is available at <https://anonymous.4open.science/r/understandingCDFSL>.

## [When are Local Queries Useful for Robust Learning?](#)

- Pascale Gourdeau · Varun Kanade · Marta Kwiatkowska · James Worrell
- abstract@[open-review](#): Distributional assumptions have been shown to be necessary for the robust learnability of concept classes when considering exact-in-the-ball robust risk and access to random examples by Gourdeau et al. (2019). In this paper, we study learning models where the learner is given more power through the use of local queries and give the first distribution-free algorithms that perform robust empirical risk minimization (ERM) for this notion of robustness. The first learning model we consider uses local membership queries (LMQ), where the learner can query the label of points near the training sample. We show that, under the uniform distribution, LMQs do not increase the robustness threshold of conjunctions and any superclass, e.g., decision lists and halfspaces. Faced with this negative result, we introduce the local equivalence query oracle, which returns whether the hypothesis and target concept agree in the perturbation region around a point in the training sample, as well as a counterexample if it exists. We show a separation result: on one hand, if the query radius  $\$\\lambda\$$  is strictly smaller than the adversary's perturbation budget  $\$\\rho\$$ , then distribution-free robust learning is impossible for a wide variety of concept classes; on the other hand, the setting  $\$\\lambda=\\rho\$$  allows us to develop robust ERM algorithms. We then bound the query complexity of these algorithms based on online learning guarantees and further improve these bounds for the special case of conjunctions. We finish by giving robust learning algorithms for halfspaces with margins on both  $\${0,1}^n\$$  and  $\$\\mathbb{R}^n\$$ .

## [A Neural Corpus Indexer for Document Retrieval](#)

- Yujing Wang · Haonan Wang · Yingyan Hou · Ziming Miao · Shibin Wu · Hao Sun · Qi Chen · Yuqing Xia · Chengmin Chi · Guoshuai Zhao · Zheng Liu · Xing Xie · Hao Sun · Weiwei Deng · Qi Zhang · Mao Yang
- abstract@[open-review](#): Current state-of-the-art document retrieval solutions mainly follow an index-retrieve paradigm, where the index is hard to be optimized for the final retrieval target. In this paper, we aim to show that an end-to-end deep neural network unifying training and indexing stages can significantly improve the recall performance of traditional methods. To this end, we propose Neural Corpus Indexer (NCI), a sequence-to-sequence network that generates relevant document identifiers directly for a designated query. To optimize the recall performance of NCI, we invent a prefix-aware weight-adaptive decoder architecture, and leverage tailored techniques including query generation, semantic document identifiers and consistency-based regularization. Empirical studies demonstrated the superiority of NCI on a commonly used academic benchmark, achieving +51.9% relative improvement on NQ320k dataset compared to the best baseline.

## [The Nature of Temporal Difference Errors in Multi-step Distributional Reinforcement Learning](#)

- Yunhao Tang · Remi Munos · Mark Rowland · Bernardo Avila Pires · Will Dabney · Marc Bellemare
- abstract@[open-review](#): We study the multi-step off-policy learning approach to distributional RL. Despite the apparent similarity between value-based RL and distributional RL, our study reveals intriguing and fundamental differences between the two cases in the multi-step setting. We identify a novel notion of path-dependent distributional TD error, which is indispensable for principled multi-step distributional RL. The distinction from the value-based case bears important implications on concepts such as backward-view algorithms. Our work provides the first theoretical guarantees on multi-step off-policy distributional RL algorithms, including results that apply to the small number of existing approaches to multi-step distributional RL. In addition, we derive a novel algorithm, Quantile Regression-Retrace, which leads to a deep RL agent QR-DQN-Retrace that shows empirical improvements over QR-DQN on the Atari-57 benchmark. Collectively, we shed light on how unique challenges in multi-step distributional RL can be addressed both in theory and practice.

## [Improving Barely Supervised Learning by Discriminating Unlabeled Samples with Super-Class](#)

- Guan Gui · Zhen Zhao · Lei Qi · Luping Zhou · Lei Wang · Yinghuan Shi
- abstract@[open-review](#): In semi-supervised learning (SSL), a common practice is to learn consistent information from unlabeled data and discriminative information from labeled data to ensure both the immutability and the separability of the classification model. Existing SSL methods suffer from failures

in barely-supervised learning (BSL), where only one or two labels per class are available, as the insufficient labels cause the discriminative information being difficult or even infeasible to learn. To bridge this gap, we investigate a simple yet effective way to leverage unlabeled samples for discriminative learning, and propose a novel discriminative information learning module to benefit model training. Specifically, we formulate the learning objective of discriminative information at the super-class level and dynamically assign different classes into different super-classes based on model performance improvement. On top of this on-the-fly process, we further propose a distribution-based loss to learn discriminative information by utilizing the similarity relationship between samples and super-classes. It encourages the unlabeled samples to stay closer to the distribution of their corresponding super-class than those of others. Such a constraint is softer than the direct assignment of pseudo labels, while the latter could be very noisy in BSL. We compare our method with state-of-the-art SSL and BSL methods through extensive experiments on standard SSL benchmarks. Our method can achieve superior results, e.g., an average accuracy of 76.76% on CIFAR-10 with merely 1 label per class.

## Discovered Policy Optimisation

- Chris Lu · Jakub Kuba · Alistair Letcher · Luke Metz · Christian Schroeder de Witt · Jakob Foerster
- abstract@[open-review](#): The last decade has been revolutionary for reinforcement learning (RL) — it can now solve complex decision and control problems. Successful RL methods were handcrafted using mathematical derivations, intuition, and experimentation. This approach has a major shortcoming—it results in specific solutions to the RL problem, rather than a protocol for discovering efficient and robust methods. In contrast, the emerging field of meta-learning provides a toolkit for automatic machine learning method optimisation, potentially addressing this flaw. However, black-box approaches which attempt to discover RL algorithms with minimal prior structure have thus far not been successful. Mirror Learning, which includes RL algorithms, such as PPO, offers a potential framework. In this paper we explore the Mirror Learning space by meta-learning a drift function. We refer to the result as Learnt Policy Optimisation (LPO). By analysing LPO we gain original insights into policy optimisation which we use to formulate a novel, closed-form RL algorithm, Discovered Policy Optimisation (DPO). Our experiments in Brax environments confirm state-of-the-art performance of LPO and DPO, as well as their transfer to unseen settings.

## Scalable Neural Video Representations with Learnable Positional Features

- Subin Kim · Sihyun Yu · Jaeho Lee · Jinwoo Shin
- abstract@[open-review](#): Succinct representation of complex signals using coordinate-based neural representations (CNRs) has seen great progress, and several recent efforts focus on extending them for handling videos. Here, the main challenge is how to (a) alleviate a compute-inefficiency in training CNRs to (b) achieve high-quality video encoding while (c) maintaining the parameter-efficiency. To meet all requirements (a), (b), and (c) simultaneously, we propose neural video representations with learnable positional features (NVP), a novel CNR by introducing "learnable positional features" that effectively amortize a video as latent codes. Specifically, we first present a CNR architecture based on designing 2D latent keyframes to learn the common video contents across each spatio-temporal axis, which dramatically improves all of those three requirements. Then, we propose to utilize existing powerful image and video codecs as a compute-/memory-efficient compression procedure of latent codes. We demonstrate the superiority of NVP on the popular UVG benchmark; compared with prior arts, NVP not only trains 2 times faster (less than 5 minutes) but also exceeds their encoding quality as  $34.00 \rightarrow 34.43$  (measured with the PSNR metric), even using  $\gg 8$  times fewer parameters. We also show intriguing properties of NVP, e.g., video inpainting, video frame interpolation, etc.

## Outlier-Robust Sparse Mean Estimation for Heavy-Tailed Distributions

- Ilias Diakonikolas · Daniel Kane · Jasper Lee · Ankit Pensia
- abstract@[open-review](#): We study the fundamental task of outlier-robust mean estimation for heavy-tailed distributions in the presence of sparsity. Specifically, given a small number of corrupted samples from a high-dimensional heavy-tailed distribution whose mean  $\mu$  is guaranteed to be sparse, the goal is to efficiently compute a hypothesis that accurately approximates  $\mu$  with high probability. Prior work had obtained efficient algorithms for robust sparse mean estimation of light-tailed distributions. In this work, we give the first sample-efficient and polynomial-time robust sparse mean estimator for heavy-tailed distributions under mild moment assumptions. Our algorithm achieves the optimal asymptotic error using a number of samples scaling logarithmically with the ambient dimension. Importantly, the sample complexity of our method is optimal as a function of the failure probability  $\tau$ , having an  $\{\text{em additive}\} \log(1/\tau)$  dependence. Our algorithm leverages the stability-based approach from the algorithmic robust statistics literature, with crucial (and necessary) adaptations required in our setting. Our analysis may be of independent interest, involving the delicate design of a (non-spectral) decomposition for positive semi-definite matrices satisfying certain sparsity properties.

## Supported Policy Optimization for Offline Reinforcement Learning

- Jialong Wu · Haixu Wu · Zihan Qiu · Jianmin Wang · Mingsheng Long
- abstract@[open-review](#): Policy constraint methods to offline reinforcement learning (RL) typically utilize parameterization or regularization that constrains the policy to perform actions within the support set of the behavior policy. The elaborate designs of parameterization methods usually intrude into the policy networks, which may bring extra inference cost and cannot take full advantage of well-established online methods. Regularization methods reduce the divergence between the learned policy and the behavior policy, which may mismatch the inherent density-based definition of support set thereby failing to avoid the out-of-distribution actions effectively. This paper presents Supported Policy OpTimization (SPOT), which is directly derived from the theoretical formalization of the density-based support constraint. SPOT adopts a VAE-based density estimator to explicitly model the support set of behavior policy and presents a simple but effective density-based regularization term, which can be plugged non-intrusively into off-the-shelf off-policy RL algorithms. SPOT achieves the state-of-the-art performance on standard benchmarks for offline RL. Benefiting from the pluggable design, offline pretrained models from SPOT can also be applied to perform online fine-tuning seamlessly.

## Subspace Recovery from Heterogeneous Data with Non-isotropic Noise

- John Duchi · Vitaly Feldman · Lunjia Hu · Kunal Talwar
- abstract@[open-review](#): Recovering linear subspaces from data is a fundamental and important task in statistics and machine learning. Motivated by heterogeneity in Federated Learning settings, we study a basic formulation of this problem: the principal component analysis (PCA), with a focus on dealing with irregular noise. Our data come from  $n$  users with user  $i$  contributing data samples from a  $d$ -dimensional distribution with mean  $\mu_i$ . Our goal is to recover the linear subspace shared by  $\mu_1, \dots, \mu_n$  using the data points from all users, where every data point from user  $i$  is formed by adding an independent mean-zero noise vector to  $\mu_i$ . If we only have one data point from every user, subspace recovery is information-theoretically impossible when the covariance matrices of the noise vectors can be non-spherical, necessitating additional restrictive assumptions in previous work. We avoid these assumptions by leveraging at least two data points from each user, which allows us to design an efficiently-computable estimator under non-spherical and user-dependent noise. We prove an upper bound for the estimation error of our estimator in general scenarios where the number of data points and amount of noise can vary across users, and prove an information-theoretic error lower bound that not only matches the upper bound up to a constant factor, but also holds even for spherical Gaussian noise. This implies that our estimator does not introduce additional estimation error (up to a constant factor) due to irregularity in the noise. We show additional results for a linear regression problem in a similar setup.

## Autoinverse: Uncertainty Aware Inversion of Neural Networks

- Navid Ansari · Hans-peter Seidel · Nima Vahidi Ferdowsi · Vahid Babaei

- abstract@[open-review](#): Neural networks are powerful surrogates for numerous forward processes. The inversion of such surrogates is extremely valuable in science and engineering. The most important property of a successful neural inverse method is the performance of its solutions when deployed in the real world, i.e., on the native forward process (and not only the learned surrogate). We propose AutoInverse, a highly automated approach for inverting neural network surrogates. Our main insight is to seek inverse solutions in the vicinity of reliable data which have been sampled from the forward process and used for training the surrogate model. AutoInverse finds such solutions by taking into account the predictive uncertainty of the surrogate and minimizing it during the inversion. Apart from high accuracy, AutoInverse enforces the feasibility of solutions, comes with embedded regularization, and is initialization free. We verify our proposed method through addressing a set of real-world problems in control, fabrication, and design.

## [CutFreq: Cut-and-Swap Frequency Components for Low-Level Vision Augmentation](#)

- Hongyang Chen · Kaisheng Ma
- abstract@[open-review](#): Low-level vision has shown great potential and importance in imaging quality and image recognition applications. However, low-level tasks suffer from limited dataset size, quality, and diversity. Data augmentation is the most effective and practical way of sample expansion, but the commonly used augmentation methods in high-level tasks have limited improvement in the low-level due to the boundary effects or the non-realistic context information. In this paper, we propose the Cut-and-Swap Frequency Components (CutFreq) method for low-level vision, which aims to preserve high-level representations with directionality and improve image synthesis quality. Observing the significant frequency domain differences between reconstructed images and real ones, in CutFreq, we propose to transform the input and real images separately in the frequency domain, then define two stages for the model training process, and finally swap the specified frequency bands respectively and inversely transform to generate augmented samples. The experimental results show the superior performance of CutFreq on five low-level vision tasks. For instance, our method improves by 0.15 dB over the SOTA method Restormer when averaged over all five image deraining datasets. We also demonstrate the effectiveness of our method in the low-data regime. The code is available in the supplementary material and will be released on GitHub.

## [Parallel Tempering With a Variational Reference](#)

- Nikola Surjanovic · Saifuddin Syed · Alexandre Bouchard-Côté · Trevor Campbell
- abstract@[open-review](#): Sampling from complex target distributions is a challenging task fundamental to Bayesian inference. Parallel tempering (PT) addresses this problem by constructing a Markov chain on the expanded state space of a sequence of distributions interpolating between the posterior distribution and a fixed reference distribution, which is typically chosen to be the prior. However, in the typical case where the prior and posterior are nearly mutually singular, PT methods are computationally prohibitive. In this work we address this challenge by constructing a generalized annealing path connecting the posterior to an adaptively tuned variational reference. The reference distribution is tuned to minimize the forward (inclusive) KL divergence to the posterior distribution using a simple, gradient-free moment-matching procedure. We show that our adaptive procedure converges to the forward KL minimizer, and that the forward KL divergence serves as a good proxy to a previously developed measure of PT performance. We also show that in the large-data limit in typical Bayesian models, the proposed method improves in performance, while traditional PT deteriorates arbitrarily. Finally, we introduce PT with two references—“one fixed, one variational”—with a novel split annealing path that ensures stable variational reference adaptation. The paper concludes with experiments that demonstrate the large empirical gains achieved by our method in a wide range of realistic Bayesian inference scenarios.

## [Performative Power](#)

- Moritz Hardt · Meena Jagadeesan · Celestine Mendler-Dünner
- abstract@[open-review](#): We introduce the notion of performative power, which measures the ability of a firm operating an algorithmic system, such as a digital content recommendation platform, to steer a population. We relate performative power to the economic theory of market power. Traditional economic concepts are well known to struggle with identifying anti-competitive patterns in digital platforms—a core challenge is the difficulty of defining the market, its participants, products, and prices. Performative power sidesteps the problem of market definition by focusing on a directly observable statistical measure instead. High performative power enables a platform to profit from steering participant behavior, whereas low performative power ensures that learning from historical data is close to optimal. Our first general result shows that under low performative power, a firm cannot do better than standard supervised learning on observed data. We draw an analogy with a firm being a price-taker, an economic condition that arises under perfect competition in classical market models. We then contrast this with a market where performative power is concentrated and show that the equilibrium state can differ significantly. We go on to study performative power in a concrete setting of strategic classification where participants can switch between competing firms. We show that monopolies maximize performative power and disutility for the participant, while competition and outside options decrease performative power. We end on a discussion of connections to measures of market power in economics and of the relationship with ongoing antitrust debates.

## [SGAM: Building a Virtual 3D World through Simultaneous Generation and Mapping](#)

- Yuan Shen · Wei-Chiu Ma · Shenlong Wang
- abstract@[open-review](#): We present simultaneous generation and mapping (SGAM), a novel 3D scene generation algorithm. Our goal is to produce a realistic, globally consistent 3D world on a large scale. Achieving this goal is challenging and goes beyond the capacities of existing 3D generation methods, which fail to scale up to create large scenes. Video generation and view synthesis methods cannot produce a globally consistent 3D scene structure. Towards tackling the challenges, we take a hybrid approach that integrates generative sensor modeling with 3D reconstruction. Our proposed approach is an autoregressive generative framework that simultaneously generates sensor data at novel viewpoints and builds a 3D map at each timestamp. Given an arbitrary camera trajectory, our method repeatedly applies this generation-and-mapping process for thousands of steps, allowing us to create a gigantic virtual world. Our model can be trained from RGB-D sequences without having access to the complete 3D scene structure. The generated scenes are readily compatible with various interactive environments and rendering engines. Building upon the CLEVR dataset, we propose a large-scale 3D scene generation benchmark and demonstrate ours can generate consistent, realistic, and geometrically-plausible scenes that compare favorably to existing view synthesis methods.

## [Fixed-Distance Hamiltonian Monte Carlo](#)

- Hadi Mohasel Afshar · Sally Cripps
- abstract@[open-review](#): We propose a variation of the Hamiltonian Monte Carlo sampling (HMC) where the equations of motion are simulated for a fixed traversed distance rather than the conventional fixed simulation time. This new mechanism tends to generate proposals that have higher target probability values. The momentum distribution that is naturally joint with our Fixed-Distance HMC (FDHMC), and keeps the proposal acceptance probability close to 1, is not Gaussian and generates momentums that have a higher expected magnitude. This translates into a reduced correlation between the successive MCMC states and according to our experimental results, can lead to a significant improvement in terms of the effective sample size per gradient when compared to the baseline HMC and No-U-Turn (NUTS) samplers.

## [Geo-Neus: Geometry-Consistent Neural Implicit Surfaces Learning for Multi-view Reconstruction](#)

- Qiancheng Fu · Qingshan Xu · Yew Soon Ong · Wenbing Tao
- abstract@[open-review](#): Recently, neural implicit surfaces learning by volume rendering has become popular for multi-view reconstruction. However, one key challenge remains: existing approaches lack explicit multi-view geometry constraints, hence usually fail to generate geometry consistent surface

reconstruction. To address this challenge, we propose geometry-consistent neural implicit surfaces learning for multi-view reconstruction. We theoretically analyze that there exists a gap between the volume rendering integral and point-based signed distance function (SDF) modeling. To bridge this gap, we directly locate the zero-level set of SDF networks and explicitly perform multi-view geometry optimization by leveraging the sparse geometry from structure from motion (SFM) and photometric consistency in multi-view stereo. This makes our SDF optimization unbiased and allows the multi-view geometry constraints to focus on the true surface optimization. Extensive experiments show that our proposed method achieves high-quality surface reconstruction in both complex thin structures and large smooth regions, thus outperforming the state-of-the-arts by a large margin.

## [Distributional Reinforcement Learning for Risk-Sensitive Policies](#)

- Shiau Hong Lim · ILYAS MALIK
- abstract@[open-review](#): We address the problem of learning a risk-sensitive policy based on the CVaR risk measure using distributional reinforcement learning. In particular, we show that the standard action-selection strategy when applying the distributional Bellman optimality operator can result in convergence to neither the dynamic, Markovian CVaR nor the static, non-Markovian CVaR. We propose modifications to the existing algorithms that include a new distributional Bellman operator and show that the proposed strategy greatly expands the utility of distributional RL in learning and representing CVaR-optimized policies. Our proposed approach is a simple extension of standard distributional RL algorithms and can therefore take advantage of many of the recent advances in deep RL. On both synthetic and real data, we empirically show that our proposed algorithm is able to learn better CVaR-optimized policies.

## [Neural Estimation of Submodular Functions with Applications to Differentiable Subset Selection](#)

- Abir De · Soumen Chakrabarti
- abstract@[open-review](#): Submodular functions and variants, through their ability to characterize diversity and coverage, have emerged as a key tool for data selection and summarization. Many recent approaches to learn submodular functions suffer from limited expressiveness. In this work, we propose FlexSubNet, a family of flexible neural models for both monotone and non-monotone submodular functions. To fit a latent submodular function from (set, value) observations, our method applies a concave function on modular functions in a recursive manner. We do not draw the concave function from a restricted family, but rather learn from data using a highly expressive neural network that implements a differentiable quadrature procedure. Such an expressive neural model for concave functions may be of independent interest. Next, we extend this setup to provide a novel characterization of monotone \$alpha\$-submodular functions, a recently introduced notion of approximate submodular functions. We then use this characterization to design a novel neural model for such functions. Finally, we consider learning submodular set functions under distant supervision in the form of (perimeter, high-value-subset) pairs. This yields a novel subset selection method based on an order-invariant, yet greedy sampler built around the above neural set functions. Our experiments on synthetic and real data show that FlexSubNet outperforms several baselines.

## [Exploration via Elliptical Episodic Bonuses](#)

- Mikael Henaff · Roberta Raileanu · Minqi Jiang · Tim Rocktäschel
- abstract@[open-review](#): In recent years, a number of reinforcement learning (RL) methods have been proposed to explore complex environments which differ across episodes. In this work, we show that the effectiveness of these methods critically relies on a count-based episodic term in their exploration bonus. As a result, despite their success in relatively simple, noise-free settings, these methods fall short in more realistic scenarios where the state space is vast and prone to noise. To address this limitation, we introduce Exploration via Elliptical Episodic Bonuses (E3B), a new method which extends count-based episodic bonuses to continuous state spaces and encourages an agent to explore states that are diverse under a learned embedding within each episode. The embedding is learned using an inverse dynamics model in order to capture controllable aspects of the environment. Our method sets a new state-of-the-art across 22 challenging tasks from the MiniHack suite, without requiring task-specific inductive biases. E3B also outperforms existing methods in reward-free exploration on Habitat, demonstrating that it can scale to high-dimensional pixel-based observations and realistic environments.

## [SIXO: Smoothing Inference with Twisted Objectives](#)

- Dieterich Lawson · Allan Raventós · Andrew Warrington · Scott Linderman
- abstract@[open-review](#): Sequential Monte Carlo (SMC) is an algorithm for approximate posterior inference in probabilistic state space models. Its efficacy is largely determined by two design choices: the proposal distribution and the sequence of target distributions. Recent work showed that the model and proposal distribution can be learned with variational techniques, maximizing a lower bound on the marginal likelihood. However, these methods are predicated on targeting the sequence of filtering distributions, conditioned only on the previous and current observations. We introduce SIXO, a variational method that learns a sequence of target distributions that approximate the smoothing distributions, incorporating information from all observations, jointly with the model and proposal. The key idea is to learn a backwards message that warps the filtering distributions into the smoothing distributions. We develop an efficient approach to learn the required backward message using density ratio estimation. We interleave this update with conventional updates for learning the model and proposal distribution. SIXO has both theoretical and practical advantages. It leads to provably tighter lower bounds and offers more accurate posterior inferences and parameter estimates in a variety of domains.

## [The Gyro-Structure of Some Matrix Manifolds](#)

- Xuan Son Nguyen
- abstract@[open-review](#): In this paper, we study the gyrovector space structure (gyro-structure) of matrix manifolds. Our work is motivated by the success of hyperbolic neural networks (HNNs) that have demonstrated impressive performance in a variety of applications. At the heart of HNNs is the theory of gyrovector spaces that provides a powerful tool for studying hyperbolic geometry. Here we focus on two matrix manifolds, i.e., Symmetric Positive Definite (SPD) and Grassmann manifolds, and consider connecting the Riemannian geometry of these manifolds with the basic operations, i.e., the binary operation and scalar multiplication on gyrovector spaces. Our work reveals some interesting facts about SPD and Grassmann manifolds. First, SPD matrices with an Affine-Invariant (AI) or a Log-Euclidean (LE) geometry have rich structure with strong connection to hyperbolic geometry. Second, linear subspaces, when equipped with our proposed basic operations, form what we call gyrocommutative and gyrononreductive gyrogroups. Furthermore, they share remarkable analogies with gyrovector spaces. We demonstrate the applicability of our approach for human activity understanding and question answering.

## [The Query Complexity of Cake Cutting](#)

- Simina Branzei · Noam Nisan
- abstract@[open-review](#): We consider the query complexity of cake cutting and give lower and upper bounds for computing approximately envy-free, perfect, and equitable allocations with the minimum number of cuts. The lower bounds are tight for computing connected envy-free allocations among \$n=3\$ players and for computing perfect and equitable allocations with minimum number of cuts between \$n=2\$ players. We also formalize moving knife procedures and show that a large subclass of this family, which captures all the known moving knife procedures, can be simulated efficiently with arbitrarily small error in the Robertson-Webb query model.

## [Rethinking Generalization in Few-Shot Classification](#)

- Markus Hiller · Rongkai Ma · Mehrtash Harandi · Tom Drummond
- abstract@[open-review](#): Single image-level annotations only correctly describe an often small subset of an image's content, particularly when complex real-world scenes are depicted. While this might be acceptable in many classification scenarios, it poses a significant challenge for applications where the set of classes differs significantly between training and test time. In this paper, we take a closer look at the implications in the context of few-shot learning. Splitting the input samples into patches and encoding these via the help of Vision Transformers allows us to establish semantic correspondences between local regions across images and independent of their respective class. The most informative patch embeddings for the task at hand are then determined as a function of the support set via online optimization at inference time, additionally providing visual interpretability of what matters most in the image. We build on recent advances in unsupervised training of networks via masked image modelling to overcome the lack of fine-grained labels and learn the more general statistical structure of the data while avoiding negative image-level annotation influence, aka supervision collapse. Experimental results show the competitiveness of our approach, achieving new state-of-the-art results on four popular few-shot classification benchmarks for 5-shot and 1-shot scenarios.

## P2P: Tuning Pre-trained Image Models for Point Cloud Analysis with Point-to-Pixel Prompting

- Ziyi Wang · Xumin Yu · Yongming Rao · Jie Zhou · Jiwen Lu
- abstract@[open-review](#): Nowadays, pre-training big models on large-scale datasets has become a crucial topic in deep learning. The pre-trained models with high representation ability and transferability are tuned on downstream tasks and achieve a great success in natural language processing and computer vision. However, it is non-trivial to promote such a pretraining-tuning paradigm to the 3D vision, given the limited training data that are relatively inconvenient to collect. In this paper, we propose a new perspective of leveraging pre-trained 2D knowledge in 3D domain to tackle this problem, tuning pre-trained image models with the novel Point-to-Pixel prompting for point cloud analysis. Following the principle of prompting engineering, we transform point clouds into colorful images with geometry-preserved projection and geometry-aware coloring to adapt to pre-trained image models, whose weights are kept frozen during the end-to-end optimization of point cloud analysis tasks. We conduct extensive experiments to demonstrate that cooperating with our proposed Point-to-Pixel Prompting, better pre-trained image model will lead to consistently better performance in 3D vision. Therefore, by leveraging prosperous development from image pre-training field, our proposed framework achieves competitive results with previous methods on point cloud classification and part segmentation.

## Estimating graphical models for count data with applications to single-cell gene network

- Feiyi Xiao · Junjie Tang · Huaying Fang · Ruibin Xi
- abstract@[open-review](#): Graphical models such as Gaussian graphical models have been widely applied for direct interaction inference in many different areas. In many modern applications, such as single-cell RNA sequencing (scRNA-seq) studies, the observed data are counts and often contain many small counts. Traditional graphical models for continuous data are inappropriate for network inference of count data. We consider the Poisson log-normal (PLN) graphical model for count data and the precision matrix of the latent normal distribution represents the network. We propose a two-step method PLNet to estimate the precision matrix. PLNet first estimates the latent covariance matrix using the maximum marginal likelihood estimator (MMLE) and then estimates the precision matrix by minimizing the lasso-penalized D-trace loss function. We establish the convergence rate of the MMLE of the covariance matrix and further establish the convergence rate and the sign consistency of the proposed PLNet estimator of the precision matrix in the high dimensional setting. Importantly, although the PLN model is not sub-Gaussian, we show that the PLNet estimator is consistent even if the model dimension goes to infinity exponentially as the sample size increases. The performance of PLNet is evaluated and compared with available methods using simulation and gene regulatory network analysis of real scRNA-seq data.

## Why GANs are overkill for NLP

- David Alvarez-Melis · Vikas Garg · Adam Kalai
- abstract@[open-review](#): This work offers a novel theoretical perspective on why, despite numerous attempts, adversarial approaches to generative modeling (e.g., GANs) have not been as popular for certain generation tasks, particularly sequential tasks such as Natural Language Generation, as they have in others, such as Computer Vision. In particular, on sequential data such as text, maximum-likelihood approaches are significantly more utilized than GANs. We show that, while it may seem that maximizing likelihood is inherently different than minimizing distinguishability, this distinction is largely artificial and only holds for limited models. We argue that minimizing KL-divergence (i.e., maximizing likelihood) is a more efficient approach to effectively minimizing the same distinguishability criteria that adversarial models seek to optimize. Reductions show that minimizing distinguishability can be seen as simply boosting likelihood for certain families of models including n-gram models and neural networks with a softmax output layer. To achieve a full polynomial-time reduction, a novel next-token distinguishability model is considered.

## Zero-Shot 3D Drug Design by Sketching and Generating

- Siyu Long · Yi Zhou · Xinyu Dai · Hao Zhou
- abstract@[open-review](#): Drug design is a crucial step in the drug discovery cycle. Recently, various deep learning-based methods design drugs by generating novel molecules from scratch, avoiding traversing large-scale drug libraries. However, they depend on scarce experimental data or time-consuming docking simulation, leading to overfitting issues with limited training data and slow generation speed. In this study, we propose the zero-shot drug design method DESERT (Drug dEsign by SKetching and geneRaTING). Specifically, DESERT splits the design process into two stages: sketching and generating, and bridges them with the molecular shape. The two-stage fashion enables our method to utilize the large-scale molecular database to reduce the need for experimental data and docking simulation. Experiments show that DESERT achieves a new state-of-the-art at a fast speed.

## Hub-Pathway: Transfer Learning from A Hub of Pre-trained Models

- Yang Shu · Zhangjie Cao · Ziyang Zhang · Jianmin Wang · Mingsheng Long
- abstract@[open-review](#): Transfer learning aims to leverage knowledge from pre-trained models to benefit the target task. Prior transfer learning work mainly transfers from a single model. However, with the emergence of deep models pre-trained from different resources, model hubs consisting of diverse models with various architectures, pre-trained datasets and learning paradigms are available. Directly applying single-model transfer learning methods to each model wastes the abundant knowledge of the model hub and suffers from high computational cost. In this paper, we propose a Hub-Pathway framework to enable knowledge transfer from a model hub. The framework generates data-dependent pathway weights, based on which we assign the pathway routes at the input level to decide which pre-trained models are activated and passed through, and then set the pathway aggregation at the output level to aggregate the knowledge from different models to make predictions. The proposed framework can be trained end-to-end with the target task-specific loss, where it learns to explore better pathway configurations and exploit the knowledge in pre-trained models for each target datum. We utilize a noisy pathway generator and design an exploration loss to further explore different pathways throughout the model hub. To fully exploit the knowledge in pre-trained models, each model is further trained by specific data that activate it, which ensures its performance and enhances knowledge transfer. Experiment results on computer vision and reinforcement learning tasks demonstrate that the proposed Hub-Pathway framework achieves the state-of-the-art performance for model hub transfer learning.

## Early Stage Convergence and Global Convergence of Training Mildly Parameterized Neural Networks

- Mingze Wang · Chao Ma

- abstract@[open-review](#): The convergence of GD and SGD when training mildly parameterized neural networks starting from random initialization is studied. For a broad range of models and loss functions, including the widely used square loss and cross entropy loss, we prove an early stage "convergence" result. We show that the loss is decreased by a significant amount in the early stage of the training, and this decreasing is fast. Furthermore, for exponential type loss functions, and under some assumptions on the training data, we show global convergence of GD. Instead of relying on extreme over-parameterization, our study is based on a microscopic analysis of the activation patterns for the neurons, which helps us derive gradient lower bounds. The results on activation patterns, which we call "neuron partition", help build intuitions for understanding the behavior of neural networks' training dynamics, and may be of independent interest.

## [Phase Transition from Clean Training to Adversarial Training](#)

- Yue Xing · Qifan Song · Guang Cheng
- abstract@[open-review](#): Adversarial training is one important algorithm to achieve robust machine learning models. However, numerous empirical results show a great performance degradation from clean training to adversarial training (e.g., 90% vs 67% testing accuracy on CIFAR-10 dataset), which does not match the theoretical guarantee delivered by the existing studies. Such a gap inspires us to explore the existence of phase transition phenomenon with respect to the attack strength: adversarial training is as well behaved as clean training in the small-attack regime, but there is a sharp transition from clean training to adversarial training in the large-attack regime. We validate this conjecture in linear regression models, and conduct comprehensive experiments in deep neural networks.

## [Unsupervised Causal Generative Understanding of Images](#)

- Titas Auciukevicius · Patrick Fox-Roberts · Edward Rosten · Paul Henderson
- abstract@[open-review](#): We present a novel framework for unsupervised object-centric 3D scene understanding that generalizes robustly to out-of-distribution images. To achieve this, we design a causal generative model reflecting the physical process by which an image is produced, when a camera captures a scene containing multiple objects. This model is trained to reconstruct multi-view images via a latent representation describing the shapes, colours and positions of the 3D objects they show. It explicitly represents object instances as separate neural radiance fields, placed into a 3D scene. We then propose an inference algorithm that can infer this latent representation given a single out-of-distribution image as input -- even when it shows an unseen combination of components, unseen spatial compositions or a radically new viewpoint. We conduct extensive experiments applying our approach to test datasets that have zero probability under the training distribution. These show that it accurately reconstructs a scene's geometry, segments objects and infers their positions, despite not receiving any supervision. Our approach significantly out-performs baselines that do not capture the true causal image generation process.

## [CEIP: Combining Explicit and Implicit Priors for Reinforcement Learning with Demonstrations](#)

- Kai Yan · Alex Schwing · Yu-Xiong Wang
- abstract@[open-review](#): Although reinforcement learning has found widespread use in dense reward settings, training autonomous agents with sparse rewards remains challenging. To address this difficulty, prior work has shown promising results when using not only task-specific demonstrations but also task-agnostic albeit somewhat related demonstrations. In most cases, the available demonstrations are distilled into an implicit prior, commonly represented via a single deep net. Explicit priors in the form of a database that can be queried have also been shown to lead to encouraging results. To better benefit from available demonstrations, we develop a method to Combine Explicit and Implicit Priors (CEIP). CEIP exploits multiple implicit priors in the form of normalizing flows in parallel to form a single complex prior. Moreover, CEIP uses an effective explicit retrieval and push-forward mechanism to condition the implicit priors. In three challenging environments, we find the proposed CEIP method to improve upon sophisticated state-of-the-art techniques.

## [Latent Planning via Expansive Tree Search](#)

- Robert Gieselmann · Florian T. Pokorny
- abstract@[open-review](#): Planning provides autonomous agents the capability to solve long-horizon decision-making problems by evaluating predictions of the future. Yet, many real-world settings exhibit high-dimensional state spaces along with unknown transition dynamics, rendering traditional planning approaches infeasible. The idea behind latent planning is to overcome those challenges by instead solving the decision-making task in a lower-dimensional embedding space. Most existing latent planners employ shooting or collocation-based trajectory optimization techniques which are known to fail in very long-horizon and highly non-convex settings. In this work, we formulate latent planning as search to discover paths between far distant states in high-dimensional and long-horizon goal-reaching scenarios. Inspired by classical sampling-based motion planning algorithms, we designed a method which explores into the latent space deeper by iteratively growing and optimizing a search tree directed towards undiscovered areas while being constrained to the estimated data support region. Our method, called Expansive Latent Space Trees (ELAST), relies on self-supervised training via contrastive learning to obtain (a) a latent state representation and (b) a latent transition density model. We embed ELAST into a model-predictive control scheme and demonstrate significant performance improvements compared to existing baselines given challenging visual control tasks in simulation, including the navigation for a deformable object.

## [Stochastic Adaptive Activation Function](#)

- Kyungsu Lee · Jaeseung Yang · Haeyun Lee · Jae Youn Hwang
- abstract@[open-review](#): Human neurons and neurotransmission mechanisms have been realized in a deep neural network by the theoretical implementations of activation functions. However, recent studies have reported that the threshold potential of neurons exhibited different values by locations and types of individual neurons, and the activation functions have limitations in representing it. Therefore, this study proposes a simple yet effective activation function that exhibits different thresholds and adaptive activations according to the positions of units and the contexts of inputs, and we denoted it as ASH activation function. ASH highlights informative features, which exhibit large values in the top percentiles in an input, whereas it rectifies low values. Most importantly, ASH exhibits trainable, adaptive, and context-aware properties compared to other activation functions. To validate the effectiveness and robustness of ASH, we implemented ASH into many deep learning models for various tasks, including classification, detection, segmentation, and image generation. The experimental analysis demonstrates that our activation function exhibits outstanding performance in any deep learning applications.

## [Biologically-plausible backpropagation through arbitrary timespans via local neuromodulators](#)

- Yuhan Helena Liu · Stephen Smith · Stefan Mihalas · Eric Shea-Brown · Uygar Saimbali
- abstract@[open-review](#): The spectacular successes of recurrent neural network models where key parameters are adjusted via backpropagation-based gradient descent have inspired much thought as to how biological neuronal networks might solve the corresponding synaptic credit assignment problem [1, 2, 3]. There is so far little agreement, however, as to how biological networks could implement the necessary backpropagation through time, given widely recognized constraints of biological synaptic network signaling architectures. Here, we propose that extra-synaptic diffusion of local neuromodulators such as neuropeptides may afford an effective mode of backpropagation lying within the bounds of biological plausibility. Going beyond existing temporal truncation-based gradient approximations [4, 5, 6], our approximate gradient-based update rule, ModProp, propagates credit information through arbitrary time steps. ModProp suggests that modulatory signals can act on receiving cells by convolving their eligibility traces via causal, time-invariant and synapse-type-specific filter taps. Our mathematical analysis of ModProp learning, together with simulation results on benchmark temporal

tasks, demonstrate the advantage of ModProp over existing biologically-plausible temporal credit assignment rules. These results suggest a potential neuronal mechanism for signaling credit information related to recurrent interactions over a longer time horizon. Finally, we derive an in-silico implementation of ModProp that could serve as a low-complexity and causal alternative to backpropagation through time.

## [Faster Deep Reinforcement Learning with Slower Online Network](#)

- Kavosh Asadi · Rasool Fakoor · Omer Gottesman · Taesup Kim · Michael Littman · Alexander Smola
- abstract@[open-review](#): Deep reinforcement learning algorithms often use two networks for value function optimization: an online network, and a target network that tracks the online network with some delay. Using two separate networks enables the agent to hedge against issues that arise when performing bootstrapping. In this paper we endow two popular deep reinforcement learning algorithms, namely DQN and Rainbow, with updates that incentivize the online network to remain in the vicinity of the target network. This improves the robustness of deep reinforcement learning in presence of noisy updates. The resultant agents, called DQN Pro and Rainbow Pro, exhibit significant performance improvements over their original counterparts on the Atari benchmark demonstrating the effectiveness of this simple idea in deep reinforcement learning.

## [PlasticityNet: Learning to Simulate Metal, Sand, and Snow for Optimization Time Integration](#)

- Xuan Li · Yadi Cao · Minchen Li · Yin Yang · Craig Schroeder · Chenfanfu Jiang
- abstract@[open-review](#): In this paper, we propose a neural network-based approach for learning to represent the behavior of plastic solid materials ranging from rubber and metal to sand and snow. Unlike elastic forces such as spring forces, these plastic forces do not result from the positional gradient of any potential energy, imposing great challenges on the stability and flexibility of their simulation. Our method effectively resolves this issue by learning a generalizable plastic energy whose derivative closely matches the analytical behavior of plastic forces. Our method, for the first time, enables the simulation of a wide range of arbitrary elasticity-plasticity combinations using time step-independent, unconditionally stable optimization-based time integrators. We demonstrate the efficacy of our method by learning and producing challenging 2D and 3D effects of metal, sand, and snow with complex dynamics.

## [Template based Graph Neural Network with Optimal Transport Distances](#)

- Cédric Vincent-Cuaz · Romain Flamary · Marco Corneli · Titouan Vayer · Nicolas Courty
- abstract@[open-review](#): Current Graph Neural Networks (GNN) architectures generally rely on two important components: node features embedding through message passing, and aggregation with a specialized form of pooling. The structural (or topological) information is implicitly taken into account in these two steps. We propose in this work a novel point of view, which places distances to some learnable graph templates at the core of the graph representation. This distance embedding is constructed thanks to an optimal transport distance: the Fused Gromov-Wasserstein (FGW) distance, which encodes simultaneously feature and structure dissimilarities by solving a soft graph-matching problem. We postulate that the vector of FGW distances to a set of template graphs has a strong discriminative power, which is then fed to a non-linear classifier for final predictions. Distance embedding can be seen as a new layer, and can leverage on existing message passing techniques to promote sensible feature representations. Interestingly enough, in our work the optimal set of template graphs is also learnt in an end-to-end fashion by differentiating through this layer. After describing the corresponding learning procedure, we empirically validate our claim on several synthetic and real life graph classification datasets, where our method is competitive or surpasses kernel and GNN state-of-the-art approaches. We complete our experiments by an ablation study and a sensitivity analysis to parameters.

## [Decision-based Black-box Attack Against Vision Transformers via Patch-wise Adversarial Removal](#)

- Yucheng Shi · Yahong Han · Yu-an Tan · Xiaohui Kuang
- abstract@[open-review](#): Vision transformers (ViTs) have demonstrated impressive performance and stronger adversarial robustness compared to Convolutional Neural Networks (CNNs). On the one hand, ViTs' focus on global interaction between individual patches reduces the local noise sensitivity of images. On the other hand, the neglect of noise sensitivity differences between image regions by existing decision-based attacks further compromises the efficiency of noise compression, especially for ViTs. Therefore, validating the black-box adversarial robustness of ViTs when the target model can only be queried still remains a challenging problem. In this paper, we theoretically analyze the limitations of existing decision-based attacks from the perspective of noise sensitivity difference between regions of the image, and propose a new decision-based black-box attack against ViTs, termed Patch-wise Adversarial Removal (PAR). PAR divides images into patches through a coarse-to-fine search process and compresses the noise on each patch separately. PAR records the noise magnitude and noise sensitivity of each patch and selects the patch with the highest query value for noise compression. In addition, PAR can be used as a noise initialization method for other decision-based attacks to improve the noise compression efficiency on both ViTs and CNNs without introducing additional calculations. Extensive experiments on three datasets demonstrate that PAR achieves a much lower noise magnitude with the same number of queries.

## [Sequence-to-Set Generative Models](#)

- Longtao Tang · Ying Zhou · Yu Yang
- abstract@[open-review](#): In this paper, we propose a sequence-to-set method that can transform any sequence generative model based on maximum likelihood to a set generative model where we can evaluate the utility/probability of any set. An efficient importance sampling algorithm is devised to tackle the computational challenge of learning our sequence-to-set model. We present GRU2Set, which is an instance of our sequence-to-set method and employs the famous GRU model as the sequence generative model. To further obtain permutation invariant representation of sets, we devise the SetNN model which is also an instance of the sequence-to-set model. A direct application of our models is to learn an order/set distribution from a collection of e-commerce orders, which is an essential step in many important operational decisions such as inventory arrangement for fast delivery. Based on the intuition that small-sized sets are usually easier to learn than large sets, we propose a size-bias trick that can help learn better set distributions with respect to the  $\ell_1$ -distance evaluation metric. Two e-commerce order datasets, TMALL and HKTVMALL, are used to conduct extensive experiments to show the effectiveness of our models. The experimental results demonstrate that our models can learn better set/order distributions from order data than the baselines. Moreover, no matter what model we use, applying the size-bias trick can always improve the quality of the set distribution learned from data.

## [Ensemble of Averages: Improving Model Selection and Boosting Performance in Domain Generalization](#)

- Devansh Arpit · Huan Wang · Yingbo Zhou · Caiming Xiong
- abstract@[open-review](#): In Domain Generalization (DG) settings, models trained independently on a given set of training domains have notoriously chaotic performance on distribution shifted test domains, and stochasticity in optimization (e.g. seed) plays a big role. This makes deep learning models unreliable in real world settings. We first show that this chaotic behavior exists even along the training optimization trajectory of a single model, and propose a simple model averaging protocol that both significantly boosts domain generalization and diminishes the impact of stochasticity by improving the rank correlation between the in-domain validation accuracy and out-domain test accuracy, which is crucial for reliable early stopping. Taking advantage of our observation, we show that instead of ensembling unaveraged models (that is typical in practice), ensembling moving average models (EoA) from independent runs further boosts performance. We theoretically explain the boost in performance of ensembling and model averaging by adapting the well known Bias-Variance trade-off to the domain generalization setting. On the DomainBed benchmark, when using a pre-trained ResNet-50, this ensemble of averages achieves an average of 68.0%, beating vanilla ERM (w/o averaging/ensembling) by ~4%, and when using a pre-trained RegNetY-16GF, achieves an average of 76.6%, beating vanilla ERM by ~6%.

## Neurosymbolic Deep Generative Models for Sequence Data with Relational Constraints

- Halley Young · Maxwell Du · Osbert Bastani
- abstract@[open-review](#): There has been significant recent progress designing deep generative models that generate realistic sequence data such as text or music. Nevertheless, it remains difficult to incorporate high-level structure to guide the generative process, and many such models perform well on local coherence, but less so on global coherence. We propose a novel approach for incorporating global structure in the form of relational constraints between different subcomponents of an example (e.g., lines of a poem or measures of music). Our generative model has two parts: (i) one model to generate a realistic set of relational constraints, and (ii) a second model to generate realistic data satisfying these constraints. For model (i), we propose a constrained optimization algorithm that infers the relational constraints present in the training data, and then learn a generative model based on the resulting constraint data. In our experiments, we show that our approach significantly improves over state-of-the-art in terms of capturing high-level structure in the data, while performing comparably or better in terms of low-level structure. We also show that using constrained optimization for part (ii) as well leads to increased controllability with little decrease in quality compared to pure learning-based models.

## One for All: Simultaneous Metric and Preference Learning over Multiple Users

- Gregory Canal · Blake Mason · Ramya Korlakai Vinayak · Robert Nowak
- abstract@[open-review](#): This paper investigates simultaneous preference and metric learning from a crowd of respondents. A set of items represented by  $d$ -dimensional feature vectors and paired comparisons of the form ``item  $i$  is preferable to item  $j$ '' made by each user is given. Our model jointly learns a distance metric that characterizes the crowd's general measure of item similarities along with a latent ideal point for each user reflecting their individual preferences. This model has the flexibility to capture individual preferences, while enjoying a metric learning sample cost that is amortized over the crowd. We first study this problem in a noiseless, continuous response setting (i.e., responses equal to differences of item distances) to understand the fundamental limits of learning. Next, we establish prediction error guarantees for noisy, binary measurements such as may be collected from human respondents, and show how the sample complexity improves when the underlying metric is low-rank. Finally, we establish recovery guarantees under assumptions on the response distribution. We demonstrate the performance of our model on both simulated data and on a dataset of color preference judgements across a large number of users.

## Asymptotics of smoothed Wasserstein distances in the small noise regime

- Yunzi Ding · Jonathan Niles-Weed
- abstract@[open-review](#): We study the behavior of the Wasserstein-\$2\$ distance between discrete measures  $\mu$  and  $\nu$  in  $\mathbb{R}^d$  when both measures are smoothed by small amounts of Gaussian noise. This procedure, known as Gaussian-smoothed optimal transport, has recently attracted attention as a statistically attractive alternative to the unregularized Wasserstein distance. We give precise bounds on the approximation properties of this proposal in the small noise regime, and establish the existence of a phase transition: we show that, if the optimal transport plan from  $\mu$  to  $\nu$  is unique and a perfect matching, there exists a critical threshold such that the difference between  $W_2(\mu, \nu)$  and the Gaussian-smoothed OT distance  $W_2(\mu \ast \mathcal{N}(\sigma), \nu \ast \mathcal{N}(\sigma))$  scales like  $\exp(-c/\sigma^2)$  for  $\sigma$  below the threshold, and scales like  $\sigma$  above it. These results establish that for  $\sigma$  sufficiently small, the smoothed Wasserstein distance approximates the unregularized distance exponentially well.

## Towards a Unified Framework for Uncertainty-aware Nonlinear Variable Selection with Theoretical Guarantees

- Wenyi Deng · Beau Coker · Rajarshi Mukherjee · Jeremiah Liu · Brent Coull
- abstract@[open-review](#): We develop a simple and unified framework for nonlinear variable importance estimation that incorporates uncertainty in the prediction function and is compatible with a wide range of machine learning models (e.g., tree ensembles, kernel methods, neural networks, etc). In particular, for a learned nonlinear model  $f(\mathbf{x})$ , we consider quantifying the importance of an input variable  $x_j$  using the integrated partial derivative  $\Psi_j = \nabla f(\mathbf{x})^T \nabla f(\mathbf{x})^{-1} \mathbf{x}_j$ . We then (1) provide a principled approach for quantifying uncertainty in variable importance by deriving its posterior distribution, and (2) show that the approach is generalizable even to non-differentiable models such as tree ensembles. Rigorous Bayesian nonparametric theorems are derived to guarantee the posterior consistency and asymptotic uncertainty of the proposed approach. Extensive simulations and experiments on healthcare benchmark datasets confirm that the proposed algorithm outperforms existing classic and recent variable selection methods.

## Distributionally Robust Optimization via Ball Oracle Acceleration

- Yair Carmon · Danielle Hausler
- abstract@[open-review](#): We develop and analyze algorithms for distributionally robust optimization (DRO) of convex losses. In particular, we consider group-structured and bounded  $f$ -divergence uncertainty sets. Our approach relies on an accelerated method that queries a ball optimization oracle, i.e., a subroutine that minimizes the objective within a small ball around the query point. Our main contribution is efficient implementations of this oracle for DRO objectives. For DRO with  $N$  non-smooth loss functions, the resulting algorithms find an  $\epsilon$ -accurate solution with  $\tilde{O}(\epsilon^{-2/3} + \epsilon^{-2})$  first-order oracle queries to individual loss functions. Compared to existing algorithms for this problem, we improve complexity by a factor of up to  $\epsilon^{-4/3}$ .

## Unified Optimal Transport Framework for Universal Domain Adaptation

- Wanxing Chang · Ye Shi · Hoang Tuan · Jingya Wang
- abstract@[open-review](#): Universal Domain Adaptation (UniDA) aims to transfer knowledge from a source domain to a target domain without any constraints on label sets. Since both domains may hold private classes, identifying target common samples for domain alignment is an essential issue in UniDA. Most existing methods require manually specified or hand-tuned threshold values to detect common samples thus they are hard to extend to more realistic UniDA because of the diverse ratios of common classes. Moreover, they cannot recognize different categories among target-private samples as these private samples are treated as a whole. In this paper, we propose to use Optimal Transport (OT) to handle these issues under a unified framework, namely UniOT. First, an OT-based partial alignment with adaptive filling is designed to detect common classes without any predefined threshold values for realistic UniDA. It can automatically discover the intrinsic difference between common and private classes based on the statistical information of the assignment matrix obtained from OT. Second, we propose an OT-based target representation learning that encourages both global discrimination and local consistency of samples to avoid the over-reliance on the source. Notably, UniOT is the first method with the capability to automatically discover and recognize private categories in the target domain for UniDA. Accordingly, we introduce a new metric  $H^3$ -score to evaluate the performance in terms of both accuracy of common samples and clustering performance of private ones. Extensive experiments clearly demonstrate the advantages of UniOT over a wide range of state-of-the-art methods in UniDA.

## Learning Individualized Treatment Rules with Many Treatments: A Supervised Clustering Approach Using Adaptive Fusion

- Haixu Ma · Donglin Zeng · Yufeng Liu
- abstract@[open-review](#): Learning an optimal Individualized Treatment Rule (ITR) is a very important problem in precision medicine. This paper is concerned with the challenge when the number of treatment arms is large, and some groups of treatments in the large treatment space may work similarly

for the patients. Motivated by the recent development of supervised clustering, we propose a novel adaptive fusion based method to cluster the treatments with similar treatment effects together and estimate the optimal ITR simultaneously through a single convex optimization. The problem is formulated as balancing  $\text{loss} + \text{penalty}$  terms with a tuning parameter, which allows the entire solution path of the treatment clustering process to be clearly visualized hierarchically. For computation, we propose an efficient algorithm based on accelerated proximal gradient and further conduct a novel group-lasso based algorithm for variable selection to boost the performance. Moreover, we demonstrate the theoretical guarantee of recovering the underlying true clustering structure of the treatments for our method. Finally, we demonstrate the superior performance of our method via both simulations and a real data application on cancer treatment, which may assist the decision making process for doctors.

## [Tensor Wheel Decomposition and Its Tensor Completion Application](#)

- Zhong-Cheng Wu · Ting-Zhu Huang · Liang-Jian Deng · Hong-Xia Dou · Deyu Meng
- abstract@[open-review](#): Recently, tensor network (TN) decompositions have gained prominence in computer vision and contributed promising results to high-order data recovery tasks. However, current TN models are rather being developed towards more intricate structures to pursue incremental improvements, which instead leads to a dramatic increase in rank numbers, thus encountering laborious hyper-parameter selection, especially for higher-order cases. In this paper, we propose a novel TN decomposition, dubbed tensor wheel (TW) decomposition, in which a high-order tensor is represented by a set of latent factors mapped into a specific wheel topology. Such decomposition is constructed starting from analyzing the graph structure, aiming to more accurately characterize the complex interactions inside objectives while maintaining a lower hyper-parameter scale, theoretically alleviating the above deficiencies. Furthermore, to investigate the potentiality of TW decomposition, we provide its one numerical application, i.e., tensor completion (TC), yet develop an efficient proximal alternating minimization-based solving algorithm with guaranteed convergence. Experimental results elaborate that the proposed method is significantly superior to other tensor decomposition-based state-of-the-art methods on synthetic and real-world data, implying the merits of TW decomposition. The code is available at: [https://github.com/zhongchengwu/code\\_TWDec](https://github.com/zhongchengwu/code_TWDec).

## [Rethinking Value Function Learning for Generalization in Reinforcement Learning](#)

- Seungyong Moon · JunYeong Lee · Hyun Oh Song
- abstract@[open-review](#): We focus on the problem of training RL agents on multiple training environments to improve the generalization performance and sample efficiency. In prior methods, policy and value function are separately optimized with the goal of avoiding interference and obtaining a more accurate value function using a separate network architecture. We identify that the value function is more prone to overfitting training data and an appropriate penalization of the value function is required for better training and test performance in the multi-environment setting. To this end, we introduce Delayed-Critic Policy Gradient (DCPG), which implicitly penalizes the value estimates by training the value function less frequently with more training data compared to the policy. Furthermore, we propose a simple self-supervised task that implicitly learns the forward and inverse dynamics of environments using a single discriminator, which can be jointly optimized with the value function. Compared to the prior methods, our proposed algorithms significantly improve generalization performance and sample efficiency in the Procgen Benchmark.

## [Video PreTraining \(VPT\): Learning to Act by Watching Unlabeled Online Videos](#)

- Bowen Baker · Ilge Akkaya · Peter Zhokov · Joost Huizinga · Jie Tang · Adrien Ecoffet · Brandon Houghton · Raul Sampedro · Jeff Clune
- abstract@[open-review](#): Pretraining on noisy, internet-scale datasets has been heavily studied as a technique for training models with broad, general capabilities for text, images, and other modalities. However, for many sequential decision domains such as robotics, video games, and computer use, publicly available data does not contain the labels required to train behavioral priors in the same way. We extend the internet-scale pretraining paradigm to sequential decision domains through semi-supervised imitation learning wherein agents learn to act by watching online unlabeled videos. Specifically, we show that with a small amount of labeled data we can train an inverse dynamics model accurate enough to label a huge unlabeled source of online data -- here, online videos of people playing Minecraft -- from which we can then train a general behavioral prior. Despite using the native human interface (mouse and keyboard at 20Hz), we show that this behavioral prior has nontrivial zero-shot capabilities and that it can be fine-tuned, with both imitation learning and reinforcement learning, to hard-exploration tasks that are impossible to learn from scratch via reinforcement learning. For many tasks our models exhibit human-level performance, and we are the first to report computer agents that can craft diamond tools, which can take proficient humans upwards of 20 minutes (24,000 environment actions) of gameplay to accomplish.

## [Transferring Fairness under Distribution Shifts via Fair Consistency Regularization](#)

- Bang An · Zora Che · Mucong Ding · Furong Huang
- abstract@[open-review](#): The increasing reliance on ML models in high-stakes tasks has raised a major concern on fairness violations. Although there has been a surge of work that improves algorithmic fairness, most of them are under the assumption of an identical training and test distribution. In many real-world applications, however, such an assumption is often violated as previously trained fair models are often deployed in a different environment, and the fairness of such models has been observed to collapse. In this paper, we study how to transfer model fairness under distribution shifts, a widespread issue in practice. We conduct a fine-grained analysis of how the fair model is affected under different types of distribution shifts, and find that domain shifts are more challenging than subpopulation shifts. Inspired by the success of self-training in transferring accuracy under domain shifts, we derive a sufficient condition for transferring group fairness. Guided by it, we propose a practical algorithm with a fair consistency regularization as the key component. A synthetic dataset benchmark, which covers all types of distribution shifts, is deployed for experimental verification of the theoretical findings. Experiments on synthetic and real datasets including image and tabular data demonstrate that our approach effectively transfers fairness and accuracy under various distribution shifts.

## [SPD: Synergy Pattern Diversifying Oriented Unsupervised Multi-agent Reinforcement Learning](#)

- Yuhang Jiang · Jianzhun Shao · Shuncheng He · Hongchang Zhang · Xiangyang Ji
- abstract@[open-review](#): Reinforcement learning typically relies heavily on a well-designed reward signal, which gets more challenging in cooperative multi-agent reinforcement learning. Alternatively, unsupervised reinforcement learning (URL) has delivered on its promise in the recent past to learn useful skills and explore the environment without external supervised signals. These approaches mainly aimed for the single agent to reach distinguishable states, insufficient for multi-agent systems due to that each agent interacts with not only the environment, but also the other agents. We propose Synergy Pattern Diversifying Oriented Unsupervised Multi-agent Reinforcement Learning (SPD) to learn generic coordination policies for agents with no extrinsic reward. Specifically, we devise the Synergy Pattern Graph (SPG), a graph depicting the relationships of agents at each time step. Furthermore, we propose an episode-wise divergence measurement to approximate the discrepancy of synergy patterns. To overcome the challenge of sparse return, we decompose the discrepancy of synergy patterns to per-time-step pseudo-reward. Empirically, we show the capacity of SPD to acquire meaningful coordination policies, such as maintaining specific formations in Multi-Agent Particle Environment and pass-and-shoot in Google Research Football. Furthermore, we demonstrate that the same instructive pretrained policy's parameters can serve as a good initialization for a series of downstream tasks' policies, achieving higher data efficiency and outperforming state-of-the-art approaches in Google Research Football.

## [Lifting the Information Ratio: An Information-Theoretic Analysis of Thompson Sampling for Contextual Bandits](#)

- Gergely Neu · Iuliia Olkhovskaia · Matteo Papini · Ludovic Schwartz
- abstract@[open-review](#): We study the Bayesian regret of the renowned Thompson Sampling algorithm in contextual bandits with binary losses and adversarially-selected contexts. We adapt the information-theoretic perspective of [cite{RvR16}](#) to the contextual setting by introducing a new concept of

information ratio based on the mutual information between the unknown model parameter and the observed loss. This allows us to bound the regret in terms of the entropy of the prior distribution through a remarkably simple proof, and with no structural assumptions on the likelihood or the prior. The extension to priors with infinite entropy only requires a Lipschitz assumption on the log-likelihood. An interesting special case is that of logistic bandits with  $d$ -dimensional parameters,  $K$  actions, and Lipschitz logits, for which we provide a  $\widetilde{O}(\sqrt{dK})$  regret upper-bound that does not depend on the smallest slope of the sigmoid link function.

## [Active Labeling: Streaming Stochastic Gradients](#)

- Vivien Cabannes · Francis Bach · Vianney Perchet · Alessandro Rudi
- abstract@[open-review](#): The workhorse of machine learning is stochastic gradient descent. To access stochastic gradients, it is common to consider iteratively input/output pairs of a training dataset. Interestingly, it appears that one does not need full supervision to access stochastic gradients, which is the main motivation of this paper. After formalizing the ``active labeling'' problem, which focuses on active learning with partial supervision, we provide a streaming technique that provably minimizes the ratio of generalization error over the number of samples. We illustrate our technique in depth for robust regression.

## [Predictive Coding beyond Gaussian Distributions](#)

- Luca Pinchetti · Tommaso Salvatori · Yordan Yordanov · Beren Millidge · Yuhang Song · Thomas Lukasiewicz
- abstract@[open-review](#): A large amount of recent research has the far-reaching goal of finding training methods for deep neural networks that can serve as alternatives to backpropagation~(BP). A prominent example is predictive coding (PC), which is a neuroscience-inspired method that performs inference on hierarchical Gaussian generative models. These methods, however, fail to keep up with modern neural networks, as they are unable to replicate the dynamics of complex layers and activation functions. In this work, we solve this problem by generalizing PC to arbitrary probability distributions, enabling the training of architectures, such as transformers, that are hard to approximate with only Gaussian assumptions. We perform three experimental analyses. First, we study the gap between our method and the standard formulation of PC on multiple toy examples. Second, we test the reconstruction quality on variational autoencoders, where our method reaches the same reconstruction quality as BP. Third, we show that our method allows us to train transformer networks and achieve performance comparable with BP on conditional language models. More broadly, this method allows neuroscience-inspired learning to be applied to multiple domains, since the internal distributions can be flexibly adapted to the data, tasks, and architectures used.

## [Provable General Function Class Representation Learning in Multitask Bandits and MDP](#)

- Rui Lu · Andrew Zhao · Simon Du · Gao Huang
- abstract@[open-review](#): While multitask representation learning has become a popular approach in reinforcement learning (RL) to boost the sample efficiency, the theoretical understanding of why and how it works is still limited. Most previous analytical works could only assume that the representation function is already known to the agent or from linear function class, since analyzing general function class representation encounters non-trivial technical obstacles such as generalization guarantee, formulation of confidence bound in abstract function space, etc. However, linear-case analysis heavily relies on the particularity of linear function class, while real-world practice usually adopts general non-linear representation functions like neural networks. This significantly reduces its applicability. In this work, we extend the analysis to general function class representations. Specifically, we consider an agent playing  $M$  contextual bandits (or MDPs) concurrently and extracting a shared representation function  $\phi$  from a specific function class  $\Phi$  using our proposed Generalized Functional Upper Confidence Bound algorithm (GFUCB). We theoretically validate the benefit of multitask representation learning within general function class for bandits and linear MDP for the first time. Lastly, we conduct experiments to demonstrate the effectiveness of our algorithm with neural net representation.

## [Attention-based Neural Cellular Automata](#)

- Mattie Tesfaldet · Derek Nowrouzezahrai · Chris Pal
- abstract@[open-review](#): Recent extensions of Cellular Automata (CA) have incorporated key ideas from modern deep learning, dramatically extending their capabilities and catalyzing a new family of Neural Cellular Automata (NCA) techniques. Inspired by Transformer-based architectures, our work presents a new class of attention-based NCAs formed using a spatially localized yet globally organized self-attention scheme. We introduce an instance of this class named Vision Transformer Cellular Automata (ViTCA). We present quantitative and qualitative results on denoising autoencoding across six benchmark datasets, comparing ViTCA to a U-Net, a U-Net-based CA baseline (UNetCA), and a Vision Transformer (ViT). When comparing across architectures configured to similar parameter complexity, ViTCA architectures yield superior performance across all benchmarks and for nearly every evaluation metric. We present an ablation study on various architectural configurations of ViTCA, an analysis of its effect on cell states, and an investigation on its inductive biases. Finally, we examine its learned representations via linear probes on its converged cell state hidden representations, yielding, on average, superior results when compared to our U-Net, ViT, and UNetCA baselines.

## [Learning a Condensed Frame for Memory-Efficient Video Class-Incremental Learning](#)

- Yixuan Pei · Zhiwu Qing · Jun CEN · Xiang Wang · Shiwei Zhang · Yaxiong Wang · Mingqian Tang · Nong Sang · Xueming Qian
- abstract@[open-review](#): Recent incremental learning for action recognition usually stores representative videos to mitigate catastrophic forgetting. However, only a few bulky videos can be stored due to the limited memory. To address this problem, we propose FrameMaker, a memory-efficient video class-incremental learning approach that learns to produce a condensed frame for each selected video. Specifically, FrameMaker is mainly composed of two crucial components: Frame Condensing and Instance-Specific Prompt. The former is to reduce the memory cost by preserving only one condensed frame instead of the whole video, while the latter aims to compensate the lost spatio-temporal details in the Frame Condensing stage. By this means, FrameMaker enables a remarkable reduction in memory but keep enough information that can be applied to following incremental tasks. Experimental results on multiple challenging benchmarks, i.e., HMDB51, UCF101 and Something-Something V2, demonstrate that FrameMaker can achieve better performance to recent advanced methods while consuming only 20% memory. Additionally, under the same memory consumption conditions, FrameMaker significantly outperforms existing state-of-the-arts by a convincing margin. The source code and models will be released when this manuscript is public.

## [Automatic differentiation of nonsmooth iterative algorithms](#)

- Jerome Bolte · Edouard Pauwels · Samuel Vaiter
- abstract@[open-review](#): Differentiation along algorithms, i.e., piggyback propagation of derivatives, is now routinely used to differentiate iterative solvers in differentiable programming. Asymptotics is well understood for many smooth problems but the nondifferentiable case is hardly considered. Is there a limiting object for nonsmooth piggyback automatic differentiation (AD)? Does it have any variational meaning and can it be used effectively in machine learning? Is there a connection with classical derivative? All these questions are addressed under appropriate contractivity conditions in the framework of conservative derivatives which has proved useful in understanding nonsmooth AD. For nonsmooth piggyback iterations, we characterize the attractor set of nonsmooth piggyback iterations as a set-valued fixed point which remains in the conservative framework. This has various consequences and in particular almost everywhere convergence of classical derivatives. Our results are illustrated on parametric convex optimization problems with forward-backward, Douglas-Rachford and Alternating Direction of Multiplier algorithms as well as the Heavy-Ball method.

## [Regret Bounds of Cooperative Thompson Sampling](#)

- Yan Chen · Qinxun Bai · Perry Dong · Maria Dimakopoulou · Wei Xu · Zhengyuan Zhou
- abstract@[open-review](#): We consider the concurrent reinforcement learning problem where  $n$  agents simultaneously learn to make decisions in the same environment by sharing experience (i.e. data) with each other. Existing works in this emerging area have empirically demonstrated that Thompson sampling (TS) based algorithms provide a particularly attractive alternative for inducing cooperation, because each agent can sample a belief environment (and compute a corresponding optimal policy) from the joint posterior computed by aggregating all agents' data, which induces diversity in exploration among agents while benefitting shared experience from all agents. However, theoretical guarantees in this area remain under-explored; in particular, no regret bound is known on TS based concurrent RL algorithms. In this paper, we fill in this gap by considering two settings. In the first, and also as a warm-up, we study the simple finite-horizon episodic RL setting, where TS is naturally adapted into the concurrent setup by having each agent sample from the current joint posterior at the beginning of each episode. We establish a  $\tilde{O}(HS\sqrt{\frac{AT}{n}})$  per-agent regret bound, where  $H$  is the horizon of the episode,  $S$  is the number of states,  $A$  is the number of actions,  $T$  is the number of episodes and  $n$  is the number of agents. This regret bound further improves to  $\tilde{O}(H\sqrt{\frac{SAT}{n}})$  under. In the second setting, we consider the infinite-horizon RL problem, where a policy is measured by its long-run average reward. Here, despite not having natural episodic breakpoints, we show that by a doubling-horizon schedule, we can adapt TS to the infinite-horizon concurrent learning setting to achieve a regret bound of  $\tilde{O}(DS\sqrt{ATn})$ , where  $D$  is the standard notion of diameter of the underlying MDP and  $T$  is the number of timesteps. Note that in both settings, the per-agent regret decreases at an optimal rate of  $\Theta(\frac{1}{\sqrt{n}})$ , which manifests the power of cooperation in concurrent RL.

## [Certifying Some Distributional Fairness with Subpopulation Decomposition](#)

- Mintong Kang · Linyi Li · Maurice Weber · Yang Liu · Ce Zhang · Bo Li
- abstract@[open-review](#): Extensive efforts have been made to understand and improve the fairness of machine learning models based on observational metrics, especially in high-stakes domains such as medical insurance, education, and hiring decisions. However, there is a lack of certified fairness considering the end-to-end performance of an ML model. In this paper, we first formulate the certified fairness of an ML model trained on a given data distribution as an optimization problem based on the model performance loss bound on a fairness constrained distribution, which is within bounded distributional distance with the training distribution. We then propose a general fairness certification framework and instantiate it for both sensitive shifting and general shifting scenarios. In particular, we propose to solve the optimization problem by decomposing the original data distribution into analytical subpopulations and proving the convexity of the subproblems to solve them. We evaluate our certified fairness on six real-world datasets and show that our certification is tight in the sensitive shifting scenario and provides non-trivial certification under general shifting. Our framework is flexible to integrate additional non-skewness constraints and we show that it provides even tighter certification under different real-world scenarios. We also compare our certified fairness bound with adapted existing distributional robustness bounds on Gaussian data and demonstrate that our method is significantly tighter.

## [End-to-End Learning to Index and Search in Large Output Spaces](#)

- Nilesh Gupta · Patrick Chen · Hsiang-Fu Yu · Cho-Jui Hsieh · Inderjit Dhillon
- abstract@[open-review](#): Extreme multi-label classification (XMC) is a popular framework for solving many real-world problems that require accurate prediction from a very large number of potential output choices. A popular approach for dealing with the large label space is to arrange the labels into a shallow tree-based index and then learn an ML model to efficiently search this index via beam search. Existing methods initialize the tree index by clustering the label space into a few mutually exclusive clusters based on pre-defined features and keep it fixed throughout the training procedure. This approach results in a sub-optimal indexing structure over the label space and limits the search performance to the quality of choices made during the initialization of the index. In this paper, we propose a novel method ELIAS which relaxes the tree-based index to a specialized weighted graph based index which is learned end-to-end with the final task objective. More specifically, ELIAS models the discrete label-to-cluster assignments in the existing tree-based index as soft learnable parameters that are learned jointly with the rest of the ML model. ELIAS achieves state-of-the-art performance on several large-scale extreme classification benchmarks with millions of labels. In particular, ELIAS can be up to 2.5% better at precision@1 and up to 4% better at recall@100 than existing XMC methods. A PyTorch implementation of ELIAS is available in the supplementary material.

## [The Stability-Efficiency Dilemma: Investigating Sequence Length Warmup for Training GPT Models](#)

- Conglong Li · Minjia Zhang · Yuxiong He
- abstract@[open-review](#): Recent works have demonstrated great success in pre-training large-scale autoregressive language models (e.g., GPT-3) on massive GPUs. To reduce the wall-clock training time, a common practice is to increase the batch size and learning rate. However, such practice is often brittle and leads to a so-called stability-efficiency dilemma: increasing the batch sizes and learning rates leads to better training efficiency but can also result in training instability, leading to poor generalization accuracy or failed runs. To better understand this phenomenon, we conduct an in-depth analysis on large-scale pre-training experiments replicating the GPT-2 model with public dataset. We find that there is a strong correlation between training instability and extreme values of gradient variance. We further identify that samples with long sequence lengths contribute to these extreme gradient variance values, especially at the beginning of the training, indicating that long sequence length can be a main source of training instability. Based on the analysis, we present a simple yet effective Sequence Length Warmup method that aims to solve the training stability-efficiency dilemma by avoiding extreme gradient variance values. Moreover, we present a lightweight tuning strategy that allows us to tune our method with just a small portion of the expensive full training. Experiments replicating GPT-2 models (117M and 1.5B) show that our approach enables stable training with 8x larger batch size and 4x larger learning rate, whereas the baseline approach struggles with training instability. To achieve the same or better zero-shot evaluation results, our method reduces the required number of training tokens and wall clock time by up to 2.2x and 3.7x, respectively. Experiments replicating GPT-3 model (125M) show that our approach enables stable training with 8x larger batch size and 40x larger learning rate, and retains 99% of the zero-shot accuracy on 11 tasks using 10x less data and 12x less time compared to the original GPT-3 training recipe, while the baseline diverges under the same settings and only retain 95% of accuracy under lower learning rate.

## [I2DFormer: Learning Image to Document Attention for Zero-Shot Image Classification](#)

- Muhammad Ferjad Naeem · Yongqin Xian · Luc V Gool · Federico Tombari
- abstract@[open-review](#): Despite the tremendous progress in zero-shot learning (ZSL), the majority of existing methods still rely on human-annotated attributes, which are difficult to annotate and scale. An unsupervised alternative is to represent each class using the word embedding associated with its semantic class name. However, word embeddings extracted from pre-trained language models do not necessarily capture visual similarities, resulting in poor zero-shot performance. In this work, we argue that online textual documents e.g., Wikipedia, contain rich visual descriptions about object classes, therefore can be used as powerful unsupervised side information for ZSL. To this end, we propose I2DFormer, a novel transformer-based ZSL framework that jointly learns to encode images and documents by aligning both modalities in a shared embedding space. In order to distill discriminative visual words from noisy documents, we introduce a new cross-modal attention module that learns fine-grained interactions between image patches and document words. Consequently, our I2DFormer not only learns highly discriminative document embeddings that capture visual similarities but also gains the ability to localize visually relevant words in image regions. Quantitatively, we demonstrate that our I2DFormer significantly outperforms previous unsupervised semantic embeddings under both zero-shot and generalized zero-shot learning settings on three public datasets. Qualitatively, we show that our method leads to highly interpretable results where document words can be grounded in the image regions.

## [Fine-tuning Language Models over Slow Networks using Activation Compression with Guarantees](#)

- Jue WANG · Binhang Yuan · Luka Rimanic · Yongjun He · Tri Dao · Beidi Chen · Christopher RÃ© · Ce Zhang
- abstract@[open-review](#): Communication compression is a crucial technique for modern distributed learning systems to alleviate their communication bottlenecks over slower networks. Despite recent intensive studies of gradient compression for data parallel-style training, compressing the activations for models trained with pipeline parallelism is still an open problem. In this paper, we propose AQ-SGD, a novel activation compression algorithm for communication-efficient pipeline parallelism training over slow networks. Different from previous efforts in activation compression, instead of compressing activation values directly, AQ-SGD compresses the changes of the activations. This allows us to show, to the best of our knowledge for the first time, that one can still achieve  $\mathcal{O}(1/\sqrt{T})$  convergence rate for non-convex objectives under activation compression, without making assumptions on gradient unbiasedness that do not hold for deep learning models with non-linear activation functions. We then show that AQ-SGD can be optimized and implemented efficiently, without additional end-to-end runtime overhead. We evaluated AQ-SGD to fine-tune language models with up to 1.5 billion parameters, compressing activations to 2-4 bits. AQ-SGD provides up to  $4.3\times$  end-to-end speed-up in slower networks, without sacrificing model quality. Moreover, we also show that AQ-SGD can be combined with state-of-the-art gradient compression algorithms to enable end-to-end communication compression: All communications between machines, including model gradients, forward activations, and backward gradients are compressed into lower precision. This provides up to  $4.9\times$  end-to-end speed-up, without sacrificing model quality.

## [On the Complexity of Adversarial Decision Making](#)

- Dylan J Foster · Alexander Rakhlin · Ayush Sekhari · Karthik Sridharan
- abstract@[open-review](#): A central problem in online learning and decision making---from bandits to reinforcement learning---is to understand what modeling assumptions lead to sample-efficient learning guarantees. With a focus on stochastic environments, a recent line of research provides general structural conditions under which sample-efficient learning is possible, but robust learning guarantees for agnostic or adversarial settings have remained elusive. We consider a general adversarial decision making framework that encompasses (structured) bandit problems with adversarial rewards and reinforcement learning problems with adversarial dynamics. Our main result is to show---via new upper and lower bounds---that the Decision-Estimation Coefficient, a complexity measure introduced by Foster et al. (2021) in the stochastic counterpart to our setting, is both necessary and sufficient for low regret in the adversarial setting. However, compared to the stochastic setting, one must apply the Decision-Estimation Coefficient to the convex hull of the class of models (or, hypotheses) under consideration. This establishes that the price of accommodating adversarial rewards or dynamics is governed by the behavior of the model class under convexification, and recovers a number of existing results---both positive and negative. En route to obtaining these guarantees, we provide new structural results that connect the Decision-Estimation Coefficient to variants of other well-known complexity measures, including the Information Ratio of Russo and Van Roy and the Exploration-by-Optimization objective of Lattimore and Győrgy.

## [Point Transformer V2: Grouped Vector Attention and Improved Sampling](#)

- Xiaoyang Wu · Yixing Lao · Li Jiang · Xihui Liu · Hengshuang Zhao
- abstract@[open-review](#): As a pioneering work exploring transformer architecture for 3D point cloud understanding, Point Transformer achieves impressive results on multiple highly competitive benchmarks. In this work, we analyze the limitations of the Point Transformer and propose our powerful and efficient Point Transformer V2 model with novel designs that overcome the limitations of previous work. In particular, we first propose group vector attention, which is more parameter-efficient and effective than the previous version of vector attention. Inheriting the advantages of both learnable weight vector and multi-head attention, we present a highly effective implementation of grouped vector attention with a novel grouped weight encoding layer. We also strengthen the position information for attention by an additional position encoding multiplier. Furthermore, we design novel and lightweight grid-based sampling methods which enable better spatial alignment for downsampling and upsampling and more efficient sampling. Extensive experiments show that our model achieves better performance than its predecessor and achieves state-of-the-art on several challenging 3D point cloud understanding benchmarks, including 3D point cloud segmentation on ScanNet v2 and S3DIS and 3D point cloud classification on ModelNet40. Our code will be made publicly available.

## [Decentralized Training of Foundation Models in Heterogeneous Environments](#)

- Binhang Yuan · Yongjun He · Tianyi Zhang · Jared Davis · Tri Dao · Beidi Chen · Percy Liang · Christopher RÃ© · Ce Zhang
- abstract@[open-review](#): Training foundation models, such as GPT-3 and PaLM, can be extremely expensive, often involving tens of thousands of GPUs running continuously for months. These models are typically trained in specialized clusters featuring fast, homogeneous interconnects and using carefully designed software systems that support both data parallelism and model/pipeline parallelism. Such dedicated clusters can be costly and difficult to obtain. Can we instead leverage the much greater amount of decentralized, heterogeneous, and lower-bandwidth interconnected compute? Previous works examining the heterogeneous, decentralized setting focus on relatively small models that can be trained in a purely data parallel manner. State-of-the-art schemes for model parallel foundation model training, such as Megatron, only consider the homogeneous data center setting. In this paper, we present the first study of training large foundation models with model parallelism in a decentralized regime over a heterogeneous network. Our key technical contribution is a scheduling algorithm that allocates different computational ``tasklets'' in the training of foundation models to a group of decentralized GPU devices connected by a slow heterogeneous network. We provide a formal cost model and further propose an efficient evolutionary algorithm to find the optimal allocation strategy. We conduct extensive experiments that represent different scenarios for learning over geo-distributed devices simulated using real-world network measurements. In the most extreme case, across 8 different cities spanning 3 continents, our approach is  $4.8\times$  faster than prior state-of-the-art training systems (Megatron).

## [Trade-off between Payoff and Model Rewards in Fair Collaborative Machine Learning](#)

- Quoc Phong Nguyen · Bryan Kian Hsiang Low · Patrick Jaillet
- abstract@[open-review](#): This paper investigates the problem of fairly trading off between payoff and model rewards in collaborative machine learning (ML) where parties aggregate their datasets together to obtain improved ML models over that of each party. Supposing parties can afford the optimal model trained on the aggregated dataset, we propose an allocation scheme that distributes the payoff fairly. Notably, the same scheme can be derived from two different approaches based on (1) desirable properties of the parties' payoffs or (2) that of the underlying payoff flows from one party to another. While the former is conceptually simpler, the latter can be used to handle the practical constraint on the budgets of parties. In particular, we propose desirable properties for achieving a fair adjustment of the payoff flows that can trade off between the model reward's performance and the payoff reward. We empirically demonstrate that our proposed scheme is a sensible solution in several scenarios of collaborative ML with different budget constraints.

## [Improving Certified Robustness via Statistical Learning with Logical Reasoning](#)

- Zhuolin Yang · Zhikuan Zhao · Boxin Wang · Jiawei Zhang · Linyi Li · Hengzhi Pei · Bojan KarlaÅ; · Ji Liu · Heng Guo · Ce Zhang · Bo Li
- abstract@[open-review](#): Intensive algorithmic efforts have been made to enable the rapid improvements of certificated robustness for complex ML models recently. However, current robustness certification methods are only able to certify under a limited perturbation radius. Given that existing pure data-driven statistical approaches have reached a bottleneck, in this paper, we propose to integrate statistical ML models with knowledge (expressed as logical rules) as a reasoning component using Markov logic networks (MLN), so as to further improve the overall certified robustness. This opens new research questions about certifying the robustness of such a paradigm, especially the reasoning component (e.g., MLN). As the first step towards understanding these questions, we first prove that the computational complexity of certifying the robustness of MLN is #P-hard. Guided by this hardness result, we then derive the first certified robustness bound for MLN by carefully analyzing different model regimes. Finally, we conduct extensive experiments on five

datasets including both high-dimensional images and natural language texts, and we show that the certified robustness with knowledge-based logical reasoning indeed significantly outperforms that of the state-of-the-arts.

## [Rethinking Image Restoration for Object Detection](#)

- Shangquan Sun · Wenqi Ren · Tao Wang · Xiaochun Cao
- abstract@[open-review](#): Although image restoration has achieved significant progress, its potential to assist object detectors in adverse imaging conditions lacks enough attention. It is reported that the existing image restoration methods cannot improve the object detector performance and sometimes even reduce the detection performance. To address the issue, we propose a targeted adversarial attack in the restoration procedure to boost object detection performance after restoration. Specifically, we present an ADAM-like adversarial attack to generate pseudo ground truth for restoration training. Resultant restored images are close to original sharp images, and at the same time, lead to better results of object detection. We conduct extensive experiments in image dehazing and low light enhancement and show the superiority of our method over conventional training and other domain adaptation and multi-task methods. The proposed pipeline can be applied to all restoration methods and detectors in both one- and two-stage.

## [Escaping Saddle Points for Effective Generalization on Class-Imbalanced Data](#)

- Harsh Rangwani · Sumukh K Aithal · Mayank Mishra · Venkatesh Babu R
- abstract@[open-review](#): Deep learning methods are primarily developed to work well on balanced data across benchmarks. However, real-world datasets exhibit imbalance of varying type and degree. Several techniques based on re-weighting and margin adjustment of loss have been proposed to enhance the performance of models on minority classes. In this work, we analyze the class-imbalanced learning problem through the lens of loss landscape. Specifically, we examine the spectral density of Hessian of class-wise loss, in which we observe that for the tail class loss landscape, the solution converges to a saddle point. Due to this, optimization methods that escape saddle points can effectively improve generalization of minority classes in class-imbalanced learning. We further theoretically and empirically demonstrate that Sharpness Aware Minimization (SAM), a recent technique that aims to converge to flat minima, can be effectively used to escape saddle points for minority classes. Using SAM results in a 6.2% increase in accuracy on the minority classes even over the state-of-the-art Vector Scaling Loss, leading to an average increase of 4% across imbalanced datasets.

## [Few-Shot Audio-Visual Learning of Environment Acoustics](#)

- Sagnik Majumder · Changan Chen · Ziad Al-Halah · Kristen Grauman
- abstract@[open-review](#): Room impulse response (RIR) functions capture how the surrounding physical environment transforms the sounds heard by a listener, with implications for various applications in AR, VR, and robotics. Whereas traditional methods to estimate RIRs assume dense geometry and/or sound measurements throughout the environment, we explore how to infer RIRs based on a sparse set of images and echoes observed in the space. Towards that goal, we introduce a transformer-based method that uses self-attention to build a rich acoustic context, then predicts RIRs of arbitrary query source-receiver locations through cross-attention. Additionally, we design a novel training objective that improves the match in the acoustic signature between the RIR predictions and the targets. In experiments using a state-of-the-art audio-visual simulator for 3D environments, we demonstrate that our method successfully generates arbitrary RIRs, outperforming state-of-the-art methods and---in a major departure from traditional methods---generalizing to novel environments in a few-shot manner.

## [Staircase Attention for Recurrent Processing of Sequences](#)

- Da JU · Stephen Roller · Sainbayar Sukhbaatar · Jason E Weston
- abstract@[open-review](#): Attention mechanisms have become a standard tool for sequence modeling tasks, in particular by stacking self-attention layers over the entire input sequence as in the Transformer architecture. In this work we introduce a novel attention procedure called staircase attention that, unlike self-attention, operates across the sequence (in time) recurrently processing the input by adding another step of processing. A step in the staircase comprises of backward tokens (encoding the sequence so far seen) and forward tokens (ingesting a new part of the sequence). Thus our model can trade off performance and compute, by increasing the amount of recurrence through time and depth. Staircase attention is shown to be able to solve tasks that involve tracking that conventional Transformers cannot, due to this recurrence. Further, it is shown to provide improved modeling power for the same size model (number of parameters) compared to self-attentive Transformers on large language modeling and dialogue tasks, yielding significant perplexity gains.

## [Meta-Reinforcement Learning with Self-Modifying Networks](#)

- Mathieu Chalvidal · Thomas Serre · Rufin VanRullen
- abstract@[open-review](#): Deep Reinforcement Learning has demonstrated the potential of neural networks tuned with gradient descent for solving complex tasks in well-delimited environments. However, these neural systems are slow learners producing specialized agents with no mechanism to continue learning beyond their training curriculum. On the contrary, biological synaptic plasticity is persistent and manifold, and has been hypothesized to play a key role in executive functions such as working memory and cognitive flexibility, potentially supporting more efficient and generic learning abilities. Inspired by this, we propose to build networks with dynamic weights, able to continually perform self-reflexive modification as a function of their current synaptic state and action-reward feedback, rather than a fixed network configuration. The resulting model, MetODS (for Meta-Optimized Dynamical Synapses) is a broadly applicable meta-reinforcement learning system able to learn efficient and powerful control rules in the agent policy space. A single layer with dynamic synapses can perform one-shot learning, generalize navigation principles to unseen environments and demonstrates a strong ability to learn adaptive motor policies, comparing favorably with previous meta-reinforcement learning approaches.

## [Nest Your Adaptive Algorithm for Parameter-Agnostic Nonconvex Minimax Optimization](#)

- Junchi YANG · Xiang Li · Niao He
- abstract@[open-review](#): Adaptive algorithms like AdaGrad and AMSGrad are successful in nonconvex optimization owing to their parameter-agnostic ability --- requiring no a priori knowledge about problem-specific parameters nor tuning of learning rates. However, when it comes to nonconvex minimax optimization, direct extensions of such adaptive optimizers without proper time-scale separation may fail to work in practice. We provide such an example proving that the simple combination of Gradient Descent Ascent (GDA) with adaptive stepsizes can diverge if the primal-dual stepsize ratio is not carefully chosen; hence, a fortiori, such adaptive extensions are not parameter-agnostic. To address the issue, we formally introduce a Nested Adaptive framework, NeAda for short, that carries an inner loop for adaptively maximizing the dual variable with controllable stopping criteria and an outer loop for adaptively minimizing the primal variable. Such mechanism can be equipped with off-the-shelf adaptive optimizers and automatically balance the progress in the primal and dual variables. Theoretically, for nonconvex-strongly-concave minimax problems, we show that NeAda with AdaGrad stepsizes can achieve the near-optimal  $\tilde{O}(\epsilon^{-2})$  and  $\tilde{O}(\epsilon^{-4})$  gradient complexities respectively in the deterministic and stochastic settings, without prior information on the problem's smoothness and strong concavity parameters. To the best of our knowledge, this is the first algorithm that simultaneously achieves near-optimal convergence rates and parameter-agnostic adaptation in the nonconvex minimax setting. Numerically, we further illustrate the robustness of the NeAda family with experiments on simple test functions and a real-world application.

## [On the Effectiveness of Fine-tuning Versus Meta-reinforcement Learning](#)

- Mandi Zhao · Pieter Abbeel · Stephen James
- abstract@[open-review](#): Intelligent agents should have the ability to leverage knowledge from previously learned tasks in order to learn new ones quickly and efficiently. Meta-learning approaches have emerged as a popular solution to achieve this. However, meta-reinforcement learning (meta-RL) algorithms have thus far been restricted to simple environments with narrow task distributions and have seen limited success. Moreover, the paradigm of pretraining followed by fine-tuning to adapt to new tasks has emerged as a simple yet effective solution in supervised learning. This calls into question the benefits of meta learning approaches also in reinforcement learning, which typically come at the cost of high complexity. We therefore investigate meta-RL approaches in a variety of vision-based benchmarks, including Procgen, RLBench, and Atari, where evaluations are made on completely novel tasks. Our findings show that when meta-learning approaches are evaluated on different tasks (rather than different variations of the same task), multi-task pretraining with fine-tuning on new tasks performs equally as well, or better, than meta-pretraining with meta test-time adaptation. This is encouraging for future research, as multi-task pretraining tends to be simpler and computationally cheaper than meta-RL. From these findings, we advocate for evaluating future meta-RL methods on more challenging tasks and including multi-task pretraining with fine-tuning as a simple, yet strong baseline.

## [Gradient Methods Provably Converge to Non-Robust Networks](#)

- Gal Vardi · Gilad Yehudai · Ohad Shamir
- abstract@[open-review](#): Despite a great deal of research, it is still unclear why neural networks are so susceptible to adversarial examples. In this work, we identify natural settings where depth-\$2\$ ReLU networks trained with gradient flow are provably non-robust (susceptible to small adversarial \$\\ell\_2\$-perturbations), even when robust networks that classify the training dataset correctly exist. Perhaps surprisingly, we show that the well-known implicit bias towards margin maximization induces bias towards non-robust networks, by proving that every network which satisfies the KKT conditions of the max-margin problem is non-robust.

## [On the Frequency-bias of Coordinate-MLPs](#)

- Sameera Ramasinghe · Lachlan E. MacDonald · Simon Lucey
- abstract@[open-review](#): We show that typical implicit regularization assumptions for deep neural networks (for regression) do not hold for coordinate-MLPs, a family of MLPs that are now ubiquitous in computer vision for representing high-frequency signals. Lack of such implicit bias disrupts smooth interpolations between training samples, and hampers generalizing across signal regions with different spectra. We investigate this behavior through a Fourier lens and uncover that as the bandwidth of a coordinate-MLP is enhanced, lower frequencies tend to get suppressed unless a suitable prior is provided explicitly. Based on these insights, we propose a simple regularization technique that can mitigate the above problem, which can be incorporated into existing networks without any architectural modifications.

## [VeriDark: A Large-Scale Benchmark for Authorship Verification on the Dark Web](#)

- Andrei Manolache · Florin Brad · Antonio Barbalau · Radu Tudor Ionescu · Marius Popescu
- abstract@[open-review](#): The Dark Web represents a hotbed for illicit activity, where users communicate on different market forums in order to exchange goods and services. Law enforcement agencies benefit from forensic tools that perform authorship analysis, in order to identify and profile users based on their textual content. However, authorship analysis has been traditionally studied using corpora featuring literary texts such as fragments from novels or fan fiction, which may not be suitable in a cybercrime context. Moreover, the few works that employ authorship analysis tools for cybercrime prevention usually employ ad-hoc experimental setups and datasets. To address these issues, we release VeriDark: a benchmark comprised of three large scale authorship verification datasets and one authorship identification dataset obtained from user activity from either Dark Web related Reddit communities or popular illicit Dark Web market forums. We evaluate competitive NLP baselines on the three datasets and perform an analysis of the predictions to better understand the limitations of such approaches. We make the datasets and baselines publicly available at <https://github.com/bit-ml/VeriDark>.

## [Fair and Efficient Allocations Without Obvious Manipulations](#)

- Alexandros Psomas · Paritosh Verma
- abstract@[open-review](#): We consider the fundamental problem of allocating a set of indivisible goods among strategic agents with additive valuation functions. It is well known that, in the absence of monetary transfers, Pareto efficient and truthful rules are dictatorial, while there is no deterministic truthful mechanism that allocates all items and achieves envy-freeness up to one item (EF1), even for the case of two agents. In this paper, we investigate the interplay of fairness and efficiency under a relaxation of truthfulness called non-obvious manipulability (NOM), recently proposed by~\citet{troyan2020obvious}. We show that this relaxation allows us to bypass the aforementioned negative results in a very strong sense. Specifically, we prove that there are deterministic and EF1 algorithms that are not obviously manipulable, and the algorithm that maximizes utilitarian social welfare (the sum of agents' utilities), which is Pareto efficient but not dictatorial, is not obviously manipulable for \$n \geq 3\$ agents (but obviously manipulable for \$n=2\$ agents). At the same time, maximizing the egalitarian social welfare (the minimum of agents' utilities) or the Nash social welfare (the product of agents' utilities) is obviously manipulable for any number of agents and items. Our main result is an approximation preserving black-box reduction from the problem of designing EF1 and NOM mechanisms to the problem of designing EF1 algorithms. En route, we prove an interesting structural result about EF1 allocations, as well as new ``best-of-both-worlds'' results (for the problem without incentives), that might be of independent interest.

## [Structure-Preserving 3D Garment Modeling with Neural Sewing Machines](#)

- Xipeng Chen · Guangrun Wang · Dizhong Zhu · Xiaodan Liang · Philip Torr · Liang Lin
- abstract@[open-review](#): 3D Garment modeling is a critical and challenging topic in the area of computer vision and graphics, with increasing attention focused on garment representation learning, garment reconstruction, and controllable garment manipulation. Whereas existing methods were constrained to model garments under specific categories or with simple topology, and failed to learn reconstructable and manipulable representations. In this paper, we propose a novel Neural Sewing Machine (NSM), a learning-based framework for structure-preserving 3D garment modeling, which is capable of modeling and learning representations for garments with diverse shapes and topologies and is successfully applied to 3D garment reconstruction and controllable manipulation. To model generic garments, we first obtain sewing pattern embedding via a unified sewing pattern encoding module as the sewing pattern can accurately describe the intrinsic structure and the topology of the 3D garment. Then we use a 3D garment decoder to decode the sewing pattern embedding into a 3D garment using the UV-position maps with masks. To preserve the intrinsic structure of the predicted 3D garment, we introduce an inner-panel structure-preserving loss, an inter-panel structure-preserving loss, and a surface-normal loss in the learning process of our framework. We evaluate NSM on the public 3D garment dataset with sewing patterns with diverse garment shapes and categories. Extensive experiments demonstrate that the proposed NSM is capable of representing 3D garments under diverse garment shapes and topologies, realistically reconstructing 3D garments from 2D images with the preserved structure, and accurately manipulating the 3D garment categories, shapes, and topologies, outperforming the state-of-the-art methods by a clear margin.

## [ProtoVAE: A Trustworthy Self-Explainable Prototypical Variational Model](#)

- Srishti Gautam · Ahcene Boubekki · Stine Hansen · Suaiaba Salahuddin · Robert Jenssen · Marina Hähne · Michael Kampffmeyer
- abstract@[open-review](#): The need for interpretable models has fostered the development of self-explainable classifiers. Prior approaches are either based on multi-stage optimization schemes, impacting the predictive performance of the model, or produce explanations that are not transparent, trustworthy or do not capture the diversity of the data. To address these shortcomings, we propose ProtoVAE, a variational autoencoder-based framework that learns class-specific prototypes in an end-to-end manner and enforces trustworthiness and diversity by regularizing the representation space and introducing an

orthonormality constraint. Finally, the model is designed to be transparent by directly incorporating the prototypes into the decision process. Extensive comparisons with previous self-explainable approaches demonstrate the superiority of ProtoVAE, highlighting its ability to generate trustworthy and diverse explanations, while not degrading predictive performance.

## [Fixing Neural Networks by Leaving the Right Past Behind](#)

- Ryutaro Tanno · Melanie F. Pradier · Aditya Nori · Yingzhen Li
- abstract@[open-review](#): Prediction failures of machine learning models often arise from deficiencies in training data, such as incorrect labels, outliers, and selection biases. However, such data points that are responsible for a given failure mode are generally not known a priori, let alone a mechanism for repairing the failure. This work draws on the Bayesian view of continual learning, and develops a generic framework for both, identifying training examples which have given rise to the target failure, and fixing the model through erasing information about them. This framework naturally allows leveraging recent advances in continual learning to this new problem of model repairment, while subsuming the existing works on influence functions and data deletion as specific instances. Experimentally, the proposed approach outperforms the baselines for both identification of detrimental training data and fixing model failures in a generalisable manner.

## [VisFIS: Improved Visual Feature Importance Supervision with Right-for-the-Right-Reason Objectives](#)

- Zhuofan Ying · Peter Hase · Mohit Bansal
- abstract@[open-review](#): Many past works aim to improve visual reasoning in models by supervising feature importance (estimated by model explanation techniques) with human annotations such as highlights of important image regions. However, recent work has shown that performance gains from feature importance (FI) supervision for Visual Question Answering (VQA) tasks persist even with random supervision, suggesting that these methods do not meaningfully align model FI with human FI. In this paper, we show that model FI supervision can meaningfully improve VQA model accuracy as well as performance on several Right-for-the-Right-Reason (RRR) metrics by optimizing for four key model objectives: (1) accurate predictions given limited but sufficient information (Sufficiency); (2) max-entropy predictions given no important information (Uncertainty); (3) invariance of predictions to changes in unimportant features (Invariance); and (4) alignment between model FI explanations and human FI explanations (Plausibility). Our best performing method, Visual Feature Importance Supervision (VisFIS), outperforms strong baselines on benchmark VQA datasets in terms of both in-distribution and out-of-distribution accuracy. While past work suggests that the mechanism for improved accuracy is through improved explanation plausibility, we show that this relationship depends crucially on explanation faithfulness (whether explanations truly represent the model's internal reasoning). Predictions are more accurate when explanations are plausible and faithful, and not when they are plausible but not faithful. Lastly, we show that, surprisingly, RRR metrics are not predictive of out-of-distribution model accuracy when controlling for a model's in-distribution accuracy, which calls into question the value of these metrics for evaluating model reasoning.

## [Text Classification with Born's Rule](#)

- Emanuele Guidotti · Alfio Ferrara
- abstract@[open-review](#): This paper presents a text classification algorithm inspired by the notion of superposition of states in quantum physics. By regarding text as a superposition of words, we derive the wave function of a document and we compute the transition probability of the document to a target class according to Born's rule. Two complementary implementations are presented. In the first one, wave functions are calculated explicitly. The second implementation embeds the classifier in a neural network architecture. Through analysis of three benchmark datasets, we illustrate several aspects of the proposed method, such as classification performance, explainability, and computational efficiency. These ideas are also applicable to non-textual data.

## [A Kernelised Stein Statistic for Assessing Implicit Generative Models](#)

- Wenkai Xu · Gesine D Reinert
- abstract@[open-review](#): Synthetic data generation has become a key ingredient for training machine learning procedures, addressing tasks such as data augmentation, analysing privacy-sensitive data, or visualising representative samples. Assessing the quality of such synthetic data generators hence has to be addressed. As (deep) generative models for synthetic data often do not admit explicit probability distributions, classical statistical procedures for assessing model goodness-of-fit may not be applicable. In this paper, we propose a principled procedure to assess the quality of a synthetic data generator. The procedure is a Kernelised Stein Discrepancy-type test which is based on a non-parametric Stein operator for the synthetic data generator of interest. This operator is estimated from samples which are obtained from the synthetic data generator and hence can be applied even when the model is only implicit. In contrast to classical testing, the sample size from the synthetic data generator can be as large as desired, while the size of the observed data that the generator aims to emulate is fixed. Experimental results on synthetic distributions and trained generative models on synthetic and real datasets illustrate that the method shows improved power performance compared to existing approaches.

## [New Lower Bounds for Private Estimation and a Generalized Fingerprinting Lemma](#)

- Gautam Kamath · Argyris Mouzakis · Vikrant Singhal
- abstract@[open-review](#): We prove new lower bounds for statistical estimation tasks under the constraint of  $(\varepsilon, \delta)$ -differential privacy. First, we provide tight lower bounds for private covariance estimation of Gaussian distributions. We show that estimating the covariance matrix in Frobenius norm requires  $\Omega(d^2)$  samples, and in spectral norm requires  $\Omega(d^{3/2})$  samples, both matching upper bounds up to logarithmic factors. We prove these bounds via our main technical contribution, a broad generalization of the fingerprinting method to exponential families. Additionally, using the private Assouad method of Acharya, Sun, and Zhang, we show a tight  $\Omega(d/\alpha^2 \varepsilon)$  lower bound for estimating the mean of a distribution with bounded covariance to  $\alpha$ -error in  $\ell_2$ -distance. Prior known lower bounds for all these problems were either polynomially weaker or held under the stricter condition of  $(\varepsilon, 0)$ -differential privacy.

## [Instability and Local Minima in GAN Training with Kernel Discriminators](#)

- Evan Becker · Parthe Pandit · Sundeep Rangan · Alyson Fletcher
- abstract@[open-review](#): Generative Adversarial Networks (GANs) are a widely-used tool for generative modeling of complex data. Despite their empirical success, the training of GANs is not fully understood due to the joint training of the generator and discriminator. This paper analyzes these joint dynamics when the true samples, as well as the generated samples, are discrete, finite sets, and the discriminator is kernel-based. A simple yet expressive framework for analyzing training called the *Isolated Points Model* is introduced. In the proposed model, the distance between true samples greatly exceeds the kernel width so that each generated point is influenced by at most one true point. The model enables precise characterization of the conditions for convergence both to good and bad minima. In particular, the analysis explains two common failure modes: (i) an approximate mode collapse and (ii) divergence. Numerical simulations are provided that predictably replicate these behaviors.

## [Deep Counterfactual Estimation with Categorical Background Variables](#)

- Edward De Brouwer

- abstract@[open-review](#): Referred to as the third rung of the causal inference ladder, counterfactual queries typically ask the "What if ?" question retrospectively. The standard approach to estimate counterfactuals resides in using a structural equation model that accurately reflects the underlying data generating process. However, such models are seldom available in practice and one usually wishes to infer them from observational data alone. Unfortunately, the correct structural equation model is in general not identifiable from the observed factual distribution. Nevertheless, in this work, we show that under the assumption that the main latent contributors to the treatment responses are categorical, the counterfactuals can be still reliably predicted. Building upon this assumption, we introduce CounterFactual Query Prediction (\method), a novel method to infer counterfactuals from continuous observations when the background variables are categorical. We show that our method significantly outperforms previously available deep-learning-based counterfactual methods, both theoretically and empirically on time series and image data. Our code is available at <https://anonymous.4open.science/r/cfqp>.

## [Beyond Adult and COMPAS: Fairness in Multi-Class Prediction](#)

- Wael Alghamdi · Hsiang Hsu · Haewon Jeong · Hao Wang · Peter Michalak · Shahab Asoodeh · Flavio Calmon
- abstract@[open-review](#): We consider the problem of producing fair probabilistic classifiers for multi-class classification tasks. We formulate this problem in terms of ``projecting'' a pre-trained (and potentially unfair) classifier onto the set of models that satisfy target group-fairness requirements. The new, projected model is given by post-processing the outputs of the pre-trained classifier by a multiplicative factor. We provide a parallelizable iterative algorithm for computing the projected classifier and derive both sample complexity and convergence guarantees. Comprehensive numerical comparisons with state-of-the-art benchmarks demonstrate that our approach maintains competitive performance in terms of accuracy-fairness trade-off curves, while achieving favorable runtime on large datasets. We also introduce an open dataset with multiple classes, multiple intersectional protected groups, and over 1M samples for benchmarking fairness interventions at scale.

## [Integral Probability Metrics PAC-Bayes Bounds](#)

- Ron Amit · Baruch Epstein · Shay Moran · Ron Meir
- abstract@[open-review](#): We present a PAC-Bayes-style generalization bound which enables the replacement of the KL-divergence with a variety of Integral Probability Metrics (IPM). We provide instances of this bound with the IPM being the total variation metric and the Wasserstein distance. A notable feature of the obtained bounds is that they naturally interpolate between classical uniform convergence bounds in the worst case (when the prior and posterior are far away from each other), and preferable bounds in better cases (when the posterior and prior are close). This illustrates the possibility of reinforcing classical generalization bounds with algorithm- and data-dependent components, thus making them more suitable to analyze algorithms that use a large hypothesis space.

## [Active Ranking without Strong Stochastic Transitivity](#)

- Hao Lou · Tao Jin · Yue Wu · Pan Xu · Quanquan Gu · Farzad Farnoud
- abstract@[open-review](#): Ranking from noisy comparisons is of great practical interest in machine learning. In this paper, we consider the problem of recovering the exact full ranking for a list of items under ranking models that do *not* assume the Strong Stochastic Transitivity property. We propose a  $\$delta$$ -correct algorithm, Probe-Rank, that actively learns the ranking of the items from noisy pairwise comparisons. We prove a sample complexity upper bound for Probe-Rank, which only depends on the preference probabilities between items that are adjacent in the true ranking. This improves upon existing sample complexity results that depend on the preference probabilities for all pairs of items. Probe-Rank thus outperforms existing methods over a large collection of instances that do not satisfy Strong Stochastic Transitivity. Thorough numerical experiments in various settings are conducted, demonstrating that Probe-Rank is significantly more sample-efficient than the state-of-the-art active ranking method.

## [Consistent Interpolating Ensembles via the Manifold-Hilbert Kernel](#)

- Yutong Wang · Clay Scott
- abstract@[open-review](#): Recent research in the theory of overparametrized learning has sought to establish generalization guarantees in the interpolating regime. Such results have been established for a few common classes of methods, but so far not for ensemble methods. We devise an ensemble classification method that simultaneously interpolates the training data, and is consistent for a broad class of data distributions. To this end, we define the manifold-Hilbert kernel for data distributed on a Riemannian manifold. We prove that kernel smoothing regression using the manifold-Hilbert kernel is weakly consistent in the setting of Devroye et al. 1998. For the sphere, we show that the manifold-Hilbert kernel can be realized as a weighted random partition kernel, which arises as an infinite ensemble of partition-based classifiers.

## [3DB: A Framework for Debugging Computer Vision Models](#)

- Guillaume Leclerc · Hadi Salman · Andrew Ilyas · Sai Vemprala · Logan Engstrom · Vibhav Vineet · Kai Xiao · Pengchuan Zhang · Shibani Santurkar · Greg Yang · Ashish Kapoor · Aleksander Madry
- abstract@[open-review](#): We introduce 3DB: an extendable, unified framework for testing and debugging vision models using photorealistic simulation. We demonstrate, through a wide range of use cases, that 3DB allows users to discover vulnerabilities in computer vision systems and gain insights into how models make decisions. 3DB captures and generalizes many robustness analyses from prior work, and enables one to study their interplay. Finally, we find that the insights generated by the system transfer to the physical world. 3DB will be released as a library alongside a set of examples and documentation. We attach 3DB to the submission.

## [PRO: Patch-level Rendering and Optimization for Infinite Visual Synthesis](#)

- Jian Liang · Chenfei Wu · Xiaowei Hu · Zhe Gan · Jianfeng Wang · Lijuan Wang · Zicheng Liu · Yuejian Fang · Nan Duan
- abstract@[open-review](#): Infinite visual synthesis aims to generate high-resolution images, long-duration videos, and even visual generation of infinite size. Some recent work tried to solve this task by first dividing data into processable patches and then training the models on them without considering the dependencies between patches. However, since they fail to model global dependencies between patches, the quality and consistency of the generation can be limited. To address this issue, we propose PRO, a patch-level \emph{render-and-optimize} strategy for infinite visual synthesis. Given a large image or a long video, PRO first splits it into non-overlapping patches and uses the ordered patch chain as a complete training instance, a rendering model autoregressively predicts each patch based on its contexts. Once a patch is predicted, it is optimized immediately and its hidden states are saved as contexts for the next \emph{render-and-optimize} process. This brings two advantages: (\$i\$) The autoregressive rendering process with information transfer between contexts provides an implicit global probabilistic distribution modeling; (\$ii\$) The timely optimization process alleviates the optimization stress of the model and helps convergence. Based on the above designs, PRO shows a strong synthesis ability on high-resolution images and long-duration videos.

## [Parametrically Retargetable Decision-Makers Tend To Seek Power](#)

- Alex Turner · Prasad Tadepalli
- abstract@[open-review](#): If capable AI agents are generally incentivized to seek power in service of the objectives we specify for them, then these systems will pose enormous risks, in addition to enormous benefits. In fully observable environments, most reward functions have an optimal policy which seeks

power by keeping options open and staying alive. However, the real world is neither fully observable, nor will agents be perfectly optimal. We consider a range of models of AI decision-making, from optimal, to random, to choices informed by learning and interacting with an environment. We discover that many decision-making functions are retargetable, and that retargetability is sufficient to cause power-seeking tendencies. Our functional criterion is simple and broad. We show that a range of qualitatively dissimilar decision-making procedures incentivize agents to seek power. We demonstrate the flexibility of our results by reasoning about learned policy incentives in Montezuma's Revenge. These results suggest a safety risk: Eventually, highly retargetable training procedures may train real-world agents which seek power over humans.

## [360-MLC: Multi-view Layout Consistency for Self-training and Hyper-parameter Tuning](#)

- Bolivar Solarte · Chin-Hsuan Wu · Yueh-Cheng Liu · Yi-Hsuan Tsai · Min Sun
- abstract@[open-review](#): We present 360-MLC, a self-training method based on multi-view layout consistency for finetuning monocular room-layout models using unlabeled 360-images only. This comes in handy in practical scenarios where a pre-trained model needs to be adapted to a new data domain without using any ground truth annotations. Our simple yet effective assumption is that multiple layout estimations in the same scene must define a consistent geometry regardless of their camera positions. Based on this idea, we leverage a pre-trained model to project estimated layout boundaries from several camera views into the 3D world coordinate. Then, we re-project them back to the spherical coordinate and build a probability function, from which we sample the pseudo-labels for self-training. To handle unconfident pseudo-labels, we evaluate the variance in the re-projected boundaries as an uncertainty value to weight each pseudo-label in our loss function during training. In addition, since ground truth annotations are not available during training nor in testing, we leverage the entropy information in multiple layout estimations as a quantitative metric to measure the geometry consistency of the scene, allowing us to evaluate any layout estimator for hyper-parameter tuning, including model selection without ground truth annotations. Experimental results show that our solution achieves favorable performance against state-of-the-art methods when self-training from three publicly available source datasets to a unique, newly labeled dataset consisting of multi-view of the same scenes.

## [Assaying Out-Of-Distribution Generalization in Transfer Learning](#)

- Florian Wenzel · Andrea Dittadi · Peter Gehler · Carl-Johann Simon-Gabriel · Max Horn · Dominik Zietlow · David Kernert · Chris Russell · Thomas Brox · Bernt Schiele · Bernhard Schäfkopf · Francesco Locatello
- abstract@[open-review](#): Since out-of-distribution generalization is a generally ill-posed problem, various proxy targets (e.g., calibration, adversarial robustness, algorithmic corruptions, invariance across shifts) were studied across different research programs resulting in different recommendations. While sharing the same aspirational goal, these approaches have never been tested under the same experimental conditions on real data. In this paper, we take a unified view of previous work, highlighting message discrepancies that we address empirically, and providing recommendations on how to measure the robustness of a model and how to improve it. To this end, we collect 172 publicly available dataset pairs for training and out-of-distribution evaluation of accuracy, calibration error, adversarial attacks, environment invariance, and synthetic corruptions. We fine-tune over 31k networks, from nine different architectures in the many- and few-shot setting. Our findings confirm that in- and out-of-distribution accuracies tend to increase jointly, but show that their relation is largely dataset-dependent, and in general more nuanced and more complex than posited by previous, smaller scale studies.

## [SatMAE: Pre-training Transformers for Temporal and Multi-Spectral Satellite Imagery](#)

- Samar Khanna · Yezhen Cong · Chenlin Meng · Patrick Liu · Erik Rozi · Yutong He · Marshall Burke · David Lobell · Stefano Ermon
- abstract@[open-review](#): Unsupervised pre-training methods for large vision models have shown to enhance performance on downstream supervised tasks. Developing similar techniques for satellite imagery presents significant opportunities as unlabelled data is plentiful and the inherent temporal and multi-spectral structure provides avenues to further improve existing pre-training strategies. In this paper, we present SatMAE, a pre-training framework for temporal or multi-spectral satellite imagery based on Masked Autoencoder (MAE). To leverage temporal information, we include a temporal embedding along with independently masking image patches across time. In addition, we demonstrate that encoding multi-spectral data as groups of bands with distinct spectral positional encodings is beneficial. Our approach yields strong improvements over previous state-of-the-art techniques, both in terms of supervised learning performance on benchmark datasets (up to  $\uparrow 7\%$ ), and transfer learning performance on downstream remote sensing tasks, including land cover classification (up to  $\uparrow 14\%$ ) and semantic segmentation.

## [DeepMed: Semiparametric Causal Mediation Analysis with Debiased Deep Learning](#)

- Siqi Xu · Lin Liu · Zhonghua Liu
- abstract@[open-review](#): Causal mediation analysis can unpack the black box of causality and is therefore a powerful tool for disentangling causal pathways in biomedical and social sciences, and also for evaluating machine learning fairness. To reduce bias for estimating Natural Direct and Indirect Effects in mediation analysis, we propose a new method called DeepMed that uses deep neural networks (DNNs) to cross-fit the infinite-dimensional nuisance functions in the efficient influence functions. We obtain novel theoretical results that our DeepMed method (1) can achieve semiparametric efficiency bound without imposing sparsity constraints on the DNN architecture and (2) can adapt to certain low-dimensional structures of the nuisance functions, significantly advancing the existing literature on DNN-based semiparametric causal inference. Extensive synthetic experiments are conducted to support our novel theoretical findings. As a proof of concept, we apply DeepMed to analyze two real datasets on machine learning fairness and reach conclusions consistent with previous findings.

## [Receding Horizon Inverse Reinforcement Learning](#)

- Yiqing Xu · Wei Gao · David Hsu
- abstract@[open-review](#): Inverse reinforcement learning (IRL) seeks to infer a cost function that explains the underlying goals and preferences of expert demonstrations. This paper presents Receding Horizon Inverse Reinforcement Learning (RHIRL), a new IRL algorithm for high-dimensional, noisy, continuous systems with black-box dynamic models. RHIRL addresses two key challenges of IRL: scalability and robustness. To handle high-dimensional continuous systems, RHIRL matches the induced optimal trajectories with expert demonstrations locally in a receding horizon manner and stitches" together the local solutions to learn the cost; it thereby avoids the curse of dimensionality". This contrasts sharply with earlier algorithms that match with expert demonstrations globally over the entire high-dimensional state space. To be robust against imperfect expert demonstrations and system control noise, RHIRL learns a state-dependent cost function ``disentangled" from system dynamics under mild conditions. Experiments on benchmark tasks show that RHIRL outperforms several leading IRL algorithms in most instances. We also prove that the cumulative error of RHIRL grows linearly with the task duration. The code is available on line.

## [Mask-based Latent Reconstruction for Reinforcement Learning](#)

- Tao Yu · Zhizheng Zhang · Cuiling Lan · Yan Lu · Zhibo Chen
- abstract@[open-review](#): For deep reinforcement learning (RL) from pixels, learning effective state representations is crucial for achieving high performance. However, in practice, limited experience and high-dimensional input prevent effective representation learning. To address this, motivated by the success of masked modeling in other research fields, we introduce mask-based reconstruction to promote state representation learning in RL. Specifically, we propose a simple yet effective self-supervised method, Mask-based Latent Reconstruction (MLR), to predict the complete state representations in the latent space from the observations with spatially and temporally masked pixels. MLR enables the better use of context information when learning state representations to make them more informative, which facilitates RL agent training. Extensive experiments show that our MLR

significantly improves the sample efficiency in RL and outperforms the state-of-the-art sample-efficient RL methods on multiple continuous and discrete control benchmarks.

## [On the Importance of Gradient Norm in PAC-Bayesian Bounds](#)

- Itai Gat · Yossi Adi · Alex Schwing · Tamir Hazan
- abstract@[open-review](#): Generalization bounds which assess the difference between the true risk and the empirical risk have been studied extensively. However, to obtain bounds, current techniques use strict assumptions such as a uniformly bounded or a Lipschitz loss function. To avoid these assumptions, in this paper, we follow an alternative approach: we relax uniform bounds assumptions by using on-average bounded loss and on-average bounded gradient norm assumptions. Following this relaxation, we propose a new generalization bound that exploits the contractivity of the log-Sobolev inequalities. These inequalities add an additional loss-gradient norm term to the generalization bound, which is intuitively a surrogate of the model complexity. We apply the proposed bound on Bayesian deep nets and empirically analyze the effect of this new loss-gradient norm term on different neural architectures.

## [Robustness to Label Noise Depends on the Shape of the Noise Distribution in Feature Space](#)

- Diane Oyen · Michal Kucer · Nicolas Hengartner · Har Simrat Singh
- abstract@[open-review](#): Machine learning classifiers have been demonstrated, both empirically and theoretically, to be robust to label noise under certain conditions --- notably the typical assumption is that label noise is independent of the features given the class label. We provide a theoretical framework that generalizes beyond this typical assumption by modeling label noise as a distribution over feature space. We show that both the scale and the \emph{shape} of the noise distribution influence the posterior likelihood; and the shape of the noise distribution has a stronger impact on classification performance if the noise is concentrated in feature space where the decision boundary can be moved. For the special case of uniform label noise (independent of features and the class label), we show that the Bayes optimal classifier for \$c\$ classes is robust to label noise until the ratio of noisy samples goes above  $\frac{c-1}{c}$  (e.g. 90% for 10 classes), which we call the \emph{tipping point}. However, for the special case of class-dependent label noise (independent of features given the class label), the tipping point can be as low as 50%. Most importantly, we show that when the noise distribution targets decision boundaries (label noise is directly dependent on feature space), classification robustness can drop off even at a small scale of noise. Even when evaluating recent label-noise mitigation methods we see reduced accuracy when label noise is dependent on features. These findings explain why machine learning often handles label noise well if the noise distribution is uniform in feature-space; yet it also points to the difficulty of overcoming label noise when it is concentrated in a region of feature space where a decision boundary can move.

## [Contrastive Neural Ratio Estimation](#)

- Benjamin K Miller · Christoph Weniger · Patrick Forr©
- abstract@[open-review](#): Likelihood-to-evidence ratio estimation is usually cast as either a binary (NRE-A) or a multiclass (NRE-B) classification task. In contrast to the binary classification framework, the current formulation of the multiclass version has an intrinsic and unknown bias term, making otherwise informative diagnostics unreliable. We propose a multiclass framework free from the bias inherent to NRE-B at optimum, leaving us in the position to run diagnostics that practitioners depend on. It also recovers NRE-A in one corner case and NRE-B in the limiting case. For fair comparison, we benchmark the behavior of all algorithms in both familiar and novel training regimes: when jointly drawn data is unlimited, when data is fixed but prior draws are unlimited, and in the commonplace fixed data and parameters setting. Our investigations reveal that the highest performing models are distant from the competitors (NRE-A, NRE-B) in hyperparameter space. We make a recommendation for hyperparameters distinct from the previous models.

## [Kernel Attractor Networks: A Unifying Framework for Memory Modeling](#)

- Georgios Iatropoulos · Johanni Brea · Wulfram Gerstner
- abstract@[open-review](#): We consider the problem of training a neural network to store a set of patterns with maximal noise robustness. A solution, in terms of optimal weights and state update rules, is derived by training each individual neuron to perform either kernel classification or interpolation with a minimum weight norm. By applying this method to feed-forward and recurrent networks, we derive optimal networks that include, as special cases, many of the hetero- and auto-associative memory models that have been proposed over the past years, such as modern Hopfield networks and Kanerva's sparse distributed memory. We generalize Kanerva's model and demonstrate a simple way to design a kernel attractor network that can store an exponential number of continuous-valued patterns with a finite basin of attraction. The framework of kernel attractor networks offers a simple and intuitive way to understand the storage capacity of previous memory models, and allows for new biological interpretations in terms of dendritic non-linearities and synaptic clustering.

## [Multi-agent Performative Prediction with Greedy Deployment and Consensus Seeking Agents](#)

- Qiang LI · Chung-Yiu Yau · Hoi-To Wai
- abstract@[open-review](#): We consider a scenario where multiple agents are learning a common decision vector from data which can be influenced by the agents' decisions. This leads to the problem of multi-agent performative prediction (Multi-PfD). In this paper, we formulate Multi-PfD as a decentralized optimization problem that minimizes a sum of loss functions, where each loss function is based on a distribution influenced by the local decision vector. We first prove the necessary and sufficient condition for the Multi-PfD problem to admit a unique multi-agent performative stable (Multi-PS) solution. We show that enforcing consensus leads to a laxer condition for existence of Multi-PS solution with respect to the distributions' sensitivities, compared to the single agent case. Then, we study a decentralized extension to the greedy deployment scheme [Mendler-Dünner et al., 2020], called the DSGD-GD scheme. We show that DSGD-GD converges to the Multi-PS solution and analyze its non asymptotic convergence rate. Numerical results validate our analysis.

## [Improved Utility Analysis of Private CountSketch](#)

- Rasmus Pagh · Mikkel Thorup
- abstract@[open-review](#): Sketching is an important tool for dealing with high-dimensional vectors that are sparse (or well-approximated by a sparse vector), especially useful in distributed, parallel, and streaming settings. It is known that sketches can be made differentially private by adding noise according to the sensitivity of the sketch, and this has been used in private analytics and federated learning settings. The post-processing property of differential privacy implies that all estimates computed from the sketch can be released within the given privacy budget. In this paper we consider the classical CountSketch, made differentially private with the Gaussian mechanism, and give an improved analysis of its estimation error. Perhaps surprisingly, the privacy-utility trade-off is essentially the best one could hope for, independent of the number of repetitions in CountSketch: The error is almost identical to the error from non-private CountSketch plus the noise needed to make the vector private in the original, high-dimensional domain.

## [CogView2: Faster and Better Text-to-Image Generation via Hierarchical Transformers](#)

- Ming Ding · Wendi Zheng · Wenyi Hong · Jie Tang
- abstract@[open-review](#): Development of transformer-based text-to-image models is impeded by its slow generation and complexity, for high-resolution images. In this work, we put forward a solution based on hierarchical transformers and local parallel autoregressive generation. We pretrain a 6B-

parameter transformer with a simple and flexible self-supervised task, a cross-modal general language model (CogLM), and fine-tune it for fast super-resolution. The new text-to-image system, CogView2, shows very competitive generation compared to concurrent state-of-the-art DALL-E-2, and naturally supports interactive text-guided editing on images.

## [Fused Orthogonal Alternating Least Squares for Tensor Clustering](#)

- Jiacheng Wang · Dan Niclae
- abstract@[open-review](#): We introduce a multi-modes tensor clustering method that implements a fused version of the alternating least squares algorithm (Fused-Orth-ALS) for simultaneous tensor factorization and clustering. The statistical convergence rates of recovery and clustering are established when the data are a noise contaminated tensor with a latent low rank CP decomposition structure. Furthermore, we show that a modified alternating least squares algorithm can provably recover the true latent low rank factorization structure when the data form an asymmetric tensor with perturbation. Clustering consistency is also established. Finally, we illustrate the accuracy and computational efficient implementation of the Fused-Orth-ALS algorithm by using both simulations and real datasets.

## [Adaptive Distribution Calibration for Few-Shot Learning with Hierarchical Optimal Transport](#)

- Dandan Guo · Long Tian · He Zhao · Mingyuan Zhou · Hongyuan Zha
- abstract@[open-review](#): Few-shot classification aims to learn a classifier to recognize unseen classes during training, where the learned model can easily become over-fitted based on the biased distribution formed by only a few training examples. A recent solution to this problem is calibrating the distribution of these few sample classes by transferring statistics from the base classes with sufficient examples, where how to decide the transfer weights from base classes to novel classes is the key. However, principled approaches for learning the transfer weights have not been carefully studied. To this end, we propose a novel distribution calibration method by learning the adaptive weight matrix between novel samples and base classes, which is built upon a hierarchical Optimal Transport (H-OT) framework. By minimizing the high-level OT distance between novel samples and base classes, we can view the learned transport plan as the adaptive weight information for transferring the statistics of base classes. The learning of the cost function between a base class and novel class in the high-level OT leads to the introduction of the low-level OT, which considers the weights of all the data samples in the base class. Experimental results on standard benchmarks demonstrate that our proposed plug-and-play model outperforms competing approaches and owns desired cross-domain generalization ability, indicating the effectiveness of the learned adaptive weights.

## [Bayesian Optimistic Optimization: Optimistic Exploration for Model-based Reinforcement Learning](#)

- Chenyang Wu · Tianci Li · Zongzhang Zhang · Yang Yu
- abstract@[open-review](#): Reinforcement learning (RL) is a general framework for modeling sequential decision making problems, at the core of which lies the dilemma of exploitation and exploration. An agent failing to explore systematically will inevitably fail to learn efficiently. Optimism in the face of uncertainty (OFU) is a conventionally successful strategy for efficient exploration. An agent following the OFU principle explores actively and efficiently. However, when applied to model-based RL, it involves specifying a confidence set of the underlying model and solving a series of nonlinear constrained optimization, which can be computationally intractable. This paper proposes an algorithm, Bayesian optimistic optimization (BOO), which adopts a dynamic weighting technique for enforcing the constraint rather than explicitly solving a constrained optimization problem. BOO is a general algorithm shown to be sample-efficient for finite spectrum RKHS. We also developed effective optimization techniques based on natural gradients and entropy regularization.

## [Training Spiking Neural Networks with Event-driven Backpropagation](#)

- YAOYU ZHU · Zhaofei Yu · Wei Fang · Xiaodong Xie · Tiejun Huang · TimothÃ©e Masquelier
- abstract@[open-review](#): Spiking Neural networks (SNNs) represent and transmit information by spatiotemporal spike patterns, which bring two major advantages: biological plausibility and suitability for ultralow-power neuromorphic implementation. Despite this, the binary firing characteristic makes training SNNs more challenging. To learn the parameters of deep SNNs in an event-driven fashion as in inference of SNNs, backpropagation with respect to spike timing is proposed. Although this event-driven learning has the advantages of lower computational cost and memory occupation, the accuracy is far below the recurrent neural network-like learning approaches. In this paper, we first analyze the commonly used temporal backpropagation training approach and prove that the sum of gradients remains unchanged between fully-connected and convolutional layers. Secondly, we show that the max pooling layer meets the above invariance rule, while the average pooling layer does not, which will suffer the gradient vanishing problem but can be revised to meet the requirement. Thirdly, we point out the gradient reverse problem in event-driven learning and propose a backward kernel that can solve this problem and keep the property of the invariable sum of gradients. The experimental results show that the proposed approach achieves the state-of-the-art performance on CIFAR10 among temporal-based training methods. Also, this is the first time that the temporal-based backpropagation approach successfully trains SNN on the CIFAR100 dataset.

## [Why Do Artificially Generated Data Help Adversarial Robustness](#)

- Yue Xing · Qifan Song · Guang Cheng
- abstract@[open-review](#): In the adversarial training framework of \cite{carmon2019unlabeled,gowal2021improving}, people use generated/real unlabeled data with pseudolabels to improve adversarial robustness. We provide statistical insights to explain why the artificially generated data improve adversarial training. In particular, we study how the attack strength and the quality of the unlabeled data affect adversarial robustness in this framework. Our results show that with a high-quality unlabeled data generator, adversarial training can benefit greatly from this framework under large attack strength, while a poor generator can still help to some extent. To make adaptions concerning the quality of generated data, we propose an algorithm that performs online adjustment to the weight between the labeled real data and the generated data, aiming to optimize the adversarial risk. Numerical studies are conducted to verify our theories and show the effectiveness of the proposed algorithm.

## [Model-based Lifelong Reinforcement Learning with Bayesian Exploration](#)

- Haotian Fu · Shangqun Yu · Michael Littman · George Konidaris
- abstract@[open-review](#): We propose a model-based lifelong reinforcement-learning approach that estimates a hierarchical Bayesian posterior distilling the common structure shared across different tasks. The learned posterior combined with a sample-based Bayesian exploration procedure increases the sample efficiency of learning across a family of related tasks. We first derive an analysis of the relationship between the sample complexity and the initialization quality of the posterior in the finite MDP setting. We next scale the approach to continuous-state domains by introducing a Variational Bayesian Lifelong Reinforcement Learning algorithm that can be combined with recent model-based deep RL methods, and that exhibits backward transfer. Experimental results on several challenging domains show that our algorithms achieve both better forward and backward transfer performance than state-of-the-art lifelong RL methods.

## [Less-forgetting Multi-lingual Fine-tuning](#)

- Yuren Mao · Yaobo Liang · Nan Duan · Haobo Wang · Kai Wang · Lu Chen · Yunjun Gao

- abstract@[open-review](#): Multi-lingual fine-tuning (MLF), which fine-tunes a multi-lingual language model (MLLM) with multiple source languages, aims to gain good zero-shot performance on target languages. In MLF, the fine-tuned model tends to fit the source languages while forgetting its cross-lingual knowledge obtained from the pre-training stage. This forgetting phenomenon degenerates the zero-shot performance of MLF, which remains under-explored. To fill this gap, this paper proposes a multi-lingual fine-tuning method, dubbed Less-forgetting Multi-lingual Fine-tuning (LF-MLF). In LF-MLF, we cast multi-lingual fine-tuning as a constrained optimization problem, where the optimization objective is to minimize forgetting, and constraints are reducing the fine-tuning loss. The proposed method has superior zero-shot performance; furthermore, it can achieve the Pareto stationarity. Extensive experiments on Named Entity Recognition, Question Answering and Natural Language Inference back up our theoretical analysis and validate the superiority of our proposals.

## [Learning Partial Equivariances From Data](#)

- David W. Romero · Suhas Lohit
- abstract@[open-review](#): Group Convolutional Neural Networks (G-CNNs) constrain learned features to respect the symmetries in the selected group, and lead to better generalization when these symmetries appear in the data. If the chosen symmetries do not occur in the data, however, equivariance leads to overly constrained models and worse performance. Frequently, transformations occurring in data can be better represented by a subset of a group than by the group as a whole, e.g., rotations in  $[-90^{\circ}, 90^{\circ}]$ . In such cases, a model that respects equivariance \textit{partially} is better suited to represent the data. In addition, relevant transformations may differ for low and high-level features. For instance, full rotation equivariance is useful to describe edge orientations in a face, but partial rotation equivariance is better suited to describe face poses relative to the camera. In other words, the optimal level of equivariance may differ per layer. In this work, we introduce \textit{Partial G-CNNs}: G-CNNs able to learn layer-wise levels of partial and full equivariance to discrete, continuous groups and combinations thereof during training. Partial G-CNNs retain full equivariance when beneficial, e.g., for rotated MNIST, but adjust it whenever it becomes harmful, e.g., for classification of 6 / 9 digits or natural images. We empirically show that partial G-CNNs pair G-CNNs when full equivariance is advantageous, and outperform them otherwise.

## [PerfectDou: Dominating DouDizhu with Perfect Information Distillation](#)

- Guan Yang · Minghuan Liu · Weijun Hong · Weinan Zhang · Fei Fang · Guangjun Zeng · Yue Lin
- abstract@[open-review](#): As a challenging multi-player card game, DouDizhu has recently drawn much attention for analyzing competition and collaboration in imperfect-information games. In this paper, we propose PerfectDou, a state-of-the-art Doudizhu AI system that summons the game, in an actor-critic framework with a proposed technique named perfect information distillation. In detail, we adopt a perfect-training-imperfection-execution framework that allows the agents to utilize the global information to guide the training of the policies as if it is a perfect information game and the trained policies can be used to play the imperfect information game during the actual gameplay. Correspondingly, we characterize card and game features for DouDizhu to represent the perfect and imperfect information. To train our system, we adopt proximal policy optimization with generalized advantage estimation in a parallel training paradigm. In experiments we show how and why PerfectDou beats all existing programs, and achieves state-of-the-art performance.

## [Scalable design of Error-Correcting Output Codes using Discrete Optimization with Graph Coloring](#)

- Samarth Gupta · Saurabh Amin
- abstract@[open-review](#): We study the problem of scalable design of Error-Correcting Output Codes (ECOC) for multi-class classification. Prior works on ECOC-based classifiers are limited to codebooks with small number of rows (classes) or columns, and do not provide optimality guarantees for the codebook design problem. We address these limitations by developing a codebook design approach based on a Mixed-Integer Quadratically Constrained Program (MIQCP). This discrete formulation is naturally suited for maximizing the error-correction capability of ECOC-based classifiers and incorporates various design criteria in a flexible manner. Our solution approach is tractable in that it incrementally increases the codebook size by adding columns to maximize the gain in error-correcting capability. In particular, we show that the maximal gain in error-correction can be upper bounded by solving a graph-coloring problem. As a result, we can efficiently generate near-optimal codebooks for very large problem instances. These codebooks provide competitive multi-class classification performance on small class datasets such as MNIST and CIFAR10. Moreover, by leveraging transfer-learned binary classifiers, we achieve better classification performance over transfer-learned multi-class CNNs on large class datasets such as CIFAR100, Caltech-101/256. Our results highlight the advantages of simple and modular ECOC-based classifiers in improving classification accuracy without the risk of overfitting.

## [Fast Neural Kernel Embeddings for General Activations](#)

- Insu Han · Amir Zandieh · Jaehoon Lee · Roman Novak · Lechao Xiao · Amin Karbasi
- abstract@[open-review](#): Infinite width limit has shed light on generalization and optimization aspects of deep learning by establishing connections between neural networks and kernel methods. Despite their importance, the utility of these kernel methods was limited in large-scale learning settings due to their (super-)quadratic runtime and memory complexities. Moreover, most prior works on neural kernels have focused on the ReLU activation, mainly due to its popularity but also due to the difficulty of computing such kernels for general activations. In this work, we overcome such difficulties by providing methods to work with general activations. First, we compile and expand the list of activation functions admitting exact dual activation expressions to compute neural kernels. When the exact computation is unknown, we present methods to effectively approximate them. We propose a fast sketching method that approximates any multi-layered Neural Network Gaussian Process (NNGP) kernel and Neural Tangent Kernel (NTK) matrices for a wide range of activation functions, going beyond the commonly analyzed ReLU activation. This is done by showing how to approximate the neural kernels using the truncated Hermite expansion of any desired activation functions. While most prior works require data points on the unit sphere, our methods do not suffer from such limitations and are applicable to any dataset of points in  $\mathbb{R}^d$ . Furthermore, we provide a subspace embedding for NNGP and NTK matrices with near input-sparsity runtime and near-optimal target dimension which applies to any \emph{homogeneous} dual activation functions with rapidly convergent Taylor expansion. Empirically, with respect to exact convolutional NTK (CNTK) computation, our method achieves  $\times 10^6$  speedup for approximate CNTK of a 5-layer Myrtle network on CIFAR-10 dataset.

## [Optimal Brain Compression: A Framework for Accurate Post-Training Quantization and Pruning](#)

- Elias Frantar · Dan Alistarh
- abstract@[open-review](#): We consider the problem of model compression for deep neural networks (DNNs) in the challenging post-training setting, in which we are given an accurate trained model, and must compress it without any retraining, based only on a small amount of calibration input data. This problem has become popular in view of the emerging software and hardware support for executing models compressed via pruning and/or quantization with speedup, and well-performing solutions have been proposed independently for both compression approaches. In this paper, we introduce a new compression framework which covers both weight pruning and quantization in a unified setting, is time- and space-efficient, and considerably improves upon the practical performance of existing post-training methods. At the technical level, our approach is based on an exact and efficient realization of the classical Optimal Brain Surgeon (OBS) framework of [LeCun, Denker, and Solla, 1990] extended to also cover weight quantization at the scale of modern DNNs, and is enabled by a series of algorithmic developments which may be of independent interest. From the practical perspective, our experimental results show that it can improve significantly upon the compression-accuracy trade-offs of existing post-training methods, and that it can even enable the accurate compound application of both pruning and quantization in a post-training setting.

## [Off-Policy Evaluation for Action-Dependent Non-stationary Environments](#)

- Yash Chandak · Shiv Shankar · Nathaniel Bastian · Bruno da Silva · Emma Brunskill · Philip Thomas
- abstract@[open-review](#): Methods for sequential decision-making are often built upon a foundational assumption that the underlying decision process is stationary. This limits the application of such methods because real-world problems are often subject to changes due to external factors (\textit{passive} non-stationarity), changes induced by interactions with the system itself (\textit{active} non-stationarity), or both (\textit{hybrid} non-stationarity). In this work, we take the first steps towards the fundamental challenge of on-policy and off-policy evaluation amidst structured changes due to active, passive, or hybrid non-stationarity. Towards this goal, we make a \textit{higher-order stationarity} assumption such that non-stationarity results in changes over time, but the way changes happen is fixed. We propose, OPEN, an algorithm that uses a double application of counterfactual reasoning and a novel importance-weighted instrument-variable regression to obtain both a lower bias and a lower variance estimate of the structure in the changes of a policy's past performances. Finally, we show promising results on how OPEN can be used to predict future performances for several domains inspired by real-world applications that exhibit non-stationarity.

## [Physics-Embedded Neural Networks: Graph Neural PDE Solvers with Mixed Boundary Conditions](#)

- Masanobu Horie · NAOTO MITSUME
- abstract@[open-review](#): Graph neural network (GNN) is a promising approach to learning and predicting physical phenomena described in boundary value problems, such as partial differential equations (PDEs) with boundary conditions. However, existing models inadequately treat boundary conditions essential for the reliable prediction of such problems. In addition, because of the locally connected nature of GNNs, it is difficult to accurately predict the state after a long time, where interaction between vertices tends to be global. We present our approach termed physics-embedded neural networks that considers boundary conditions and predicts the state after a long time using an implicit method. It is built based on an  $\mathcal{E}(n)$ -equivariant GNN, resulting in high generalization performance on various shapes. We demonstrate that our model learns flow phenomena in complex shapes and outperforms a well-optimized classical solver and a state-of-the-art machine learning model in speed-accuracy trade-off. Therefore, our model can be a useful standard for realizing reliable, fast, and accurate GNN-based PDE solvers.

## [Neural Payoff Machines: Predicting Fair and Stable Payoff Allocations Among Team Members](#)

- Daphne Cornelisse · Thomas Rood · Yoram Bachrach · Mateusz Malinowski · Tal Kachman
- abstract@[open-review](#): In many multi-agent settings, participants can form teams to achieve collective outcomes that may far surpass their individual capabilities. Measuring the relative contributions of agents and allocating them shares of the reward that promote long-lasting cooperation are difficult tasks. Cooperative game theory offers solution concepts identifying distribution schemes, such as the Shapley value, that fairly reflect the contribution of individuals to the performance of the team or the Core, which reduces the incentive of agents to abandon their team. Applications of such methods include identifying influential features and sharing the costs of joint ventures or team formation. Unfortunately, using these solutions requires tackling a computational barrier as they are hard to compute, even in restricted settings. In this work, we show how cooperative game-theoretic solutions can be distilled into a learned model by training neural networks to propose fair and stable payoff allocations. We show that our approach creates models that can generalize to games far from the training distribution and can predict solutions for more players than observed during training. An important application of our framework is Explainable AI: our approach can be used to speed-up Shapley value computations on many instances.

## [Best of Both Worlds Model Selection](#)

- Aldo Pacchiano · Christoph Dann · Claudio Gentile
- abstract@[open-review](#): We study the problem of model selection in bandit scenarios in the presence of nested policy classes, with the goal of obtaining simultaneous adversarial and stochastic (best of both worlds) high-probability regret guarantees. Our approach requires that each base learner comes with a candidate regret bound that may or may not hold, while our meta algorithm plays each base learner according to a schedule that keeps the base learner's candidate regret bounds balanced until they are detected to violate their guarantees. We develop careful mis-specification tests specifically designed to blend the above model selection criterion with the ability to leverage the (potentially benign) nature of the environment. We recover the model selection guarantees of the Corral algorithm (Agarwal et al. 2017) for adversarial environments, but with the additional benefit of achieving high probability regret bounds, specifically in the case of nested adversarial linear bandits. More importantly, our model selection results also hold simultaneously in stochastic environments under gap assumptions. These are the first theoretical results that achieve best of both world (stochastic and adversarial) guarantees while performing model selection in bandit scenarios.

## [Independence Testing for Bounded Degree Bayesian Networks](#)

- Arnab Bhattacharyya · Clément L Canonne · Qiping Yang
- abstract@[open-review](#): We study the following independence testing problem: given access to samples from a distribution  $P$  over  $\{0,1\}^n$ , decide whether  $P$  is a product distribution or whether it is  $\varepsilon$ -far in total variation distance from any product distribution. For arbitrary distributions, this problem requires  $\exp(n)$  samples. We show in this work that if  $P$  has a sparse structure, then in fact only linearly many samples are required. Specifically, if  $P$  is Markov with respect to a Bayesian network whose underlying DAG has in-degree bounded by  $d$ , then  $\tilde{\Theta}(2^{d/2} \cdot n \cdot \varepsilon^2)$  samples are necessary and sufficient for independence testing.

## [Masked Conditional Video Diffusion for Prediction, Generation, and Interpolation](#)

- Vikram Voleti · Alexia Jolicoeur-Martineau · Chris Pal
- abstract@[open-review](#): Video prediction is a challenging task. The quality of video frames from current state-of-the-art (SOTA) generative models tends to be poor and generalization beyond the training data is difficult. Furthermore, existing prediction frameworks are typically not capable of simultaneously handling other video-related tasks such as unconditional generation or interpolation. In this work, we devise a general-purpose framework called Masked Conditional Video Diffusion (MCVD) for all of these video synthesis tasks using a probabilistic conditional score-based denoising diffusion model, conditioned on past and/or future frames. We train the model in a manner where we randomly and independently mask all the past frames or all the future frames. This novel but straightforward setup allows us to train a single model that is capable of executing a broad range of video tasks, specifically: unconditional generation -- where both past and future frames are masked; forward prediction -- only future frames are masked; reverse extrapolation -- only the past frames are masked; and interpolation -- neither past nor future frames are masked. We generate videos of arbitrary lengths autoregressively in a block-wise manner. Our experiments show that this approach can generate high-quality frames for diverse types of videos. Our MCVD models are built from simple non-recurrent 2D-convolutional architectures, conditioning on blocks of frames and generating blocks of frames. Our approach yields SOTA results across standard video prediction benchmarks, with computation times for training models measured in 1-12 days using  $\leq 4$  GPUs.

## [VisCo Grids: Surface Reconstruction with Viscosity and Coarea Grids](#)

- Albert Pumarola · Artsiom Sanakoyeu · Lior Yariv · Ali Thabet · Yaron Lipman
- abstract@[open-review](#): Surface reconstruction has been seeing a lot of progress lately by utilizing Implicit Neural Representations (INRs). Despite their success, INRs often introduce hard to control inductive bias (i.e., the solution surface can exhibit unexplainable behaviours), have costly inference, and are slow to train. The goal of this work is to show that replacing neural networks with simple grid functions, along with two novel geometric priors achieve comparable results to INRs, with instant inference, and improved training times. To that end we introduce VisCo Grids: a grid-based surface reconstruction method incorporating Viscosity and Coarea priors. Intuitively, the Viscosity prior replaces the smoothness inductive bias of INRs, while the

Coarea favors a minimal area solution. Experimenting with VisCo Grids on a standard reconstruction baseline provided comparable results to the best performing INRs on this dataset.

## [Combining Implicit and Explicit Regularization for Efficient Learning in Deep Networks](#)

- Dan Zhao
- abstract@[open-review](#): Recent work on implicit regularization has focused on gradient trajectories during the optimization process in attempting to explain why deep networks favor certain kinds of solutions over others. For deep linear neural networks, it has been shown that gradient descent/flow implicit regularizes toward low-rank solutions in matrix completion/factorization tasks, similar to an accelerative pre-conditioning, whose effects become more pronounced with increased depth. In light of this, we propose an explicit penalty that mirrors the rank minimization behavior and generalization performance independently of depth, but interestingly only takes effect with Adam and some of its close variantsâ€”it outperforms many approaches in matrix completion and is robust to a wide range of parameter and data regimes. Our findings suggest that explicit regularization can play a key role together with the choice of optimization algorithm in designing different, desirable forms of regularization, and that a more nuanced understanding of this interplay may be necessary.

## [Normalizing Flows for Knockoff-free Controlled Feature Selection](#)

- Derek Hansen Â· Brian Manzo Â· Jeffrey Regier
- abstract@[open-review](#): Controlled feature selection aims to discover the features a response depends on while limiting the false discovery rate (FDR) to a predefined level. Recently, multiple deep-learning-based methods have been proposed to perform controlled feature selection through the Model-X knockoff framework. We demonstrate, however, that these methods often fail to control the FDR for two reasons. First, these methods often learn inaccurate models of features. Second, the "swap" property, which is required for knockoffs to be valid, is often not well enforced. We propose a new procedure called FlowSelect to perform controlled feature selection that does not suffer from either of these two problems. To more accurately model the features, FlowSelect uses normalizing flows, the state-of-the-art method for density estimation. Instead of enforcing the "swap" property, FlowSelect uses a novel MCMC-based procedure to calculate p-values for each feature directly. Asymptotically, FlowSelect computes valid p-values. Empirically, FlowSelect consistently controls the FDR on both synthetic and semi-synthetic benchmarks, whereas competing knockoff-based approaches do not. FlowSelect also demonstrates greater power on these benchmarks. Additionally, FlowSelect correctly infers the genetic variants associated with specific soybean traits from GWAS data.

## [The Franz-Parisi Criterion and Computational Trade-offs in High Dimensional Statistics](#)

- Afonso Bandeira Â· Ahmed El Alaoui Â· Samuel Hopkins Â· Tselil Schramm Â· Alexander Wein Â· Ilias Zadik
- abstract@[open-review](#): Many high-dimensional statistical inference problems are believed to possess inherent computational hardness. Various frameworks have been proposed to give rigorous evidence for such hardness, including lower bounds against restricted models of computation (such as low-degree functions), as well as methods rooted in statistical physics that are based on free energy landscapes. This paper aims to make a rigorous connection between the seemingly different low-degree and free-energy based approaches. We define a free-energy based criterion for hardness and formally connect it to the well-established notion of low-degree hardness for a broad class of statistical problems, namely all Gaussian additive models and certain models with a sparse planted signal. By leveraging these rigorous connections we are able to: establish that for Gaussian additive models the "algebraic" notion of low-degree hardness implies failure of "geometric" local MCMC algorithms, and provide new low-degree lower bounds for sparse linear regression which seem difficult to prove directly. These results provide both conceptual insights into the connections between different notions of hardness, as well as concrete technical tools such as new methods for proving low-degree lower bounds.

## [Mean Estimation in High-Dimensional Binary Markov Gaussian Mixture Models](#)

- Yihan Zhang Â· Nir Weinberger
- abstract@[open-review](#): We consider a high-dimensional mean estimation problem over a binary hidden Markov model, which illuminates the interplay between memory in data, sample size, dimension, and signal strength in statistical inference. In this model, an estimator observes  $n$  samples of a  $d$ -dimensional parameter vector  $\theta \in \mathbb{R}^d$ , multiplied by a random sign  $S_i \in \{-1, 1\}$ , and corrupted by isotropic standard Gaussian noise. The sequence of signs  $(S_i)_{i \in [n]}$  is drawn from a stationary homogeneous Markov chain with flip probability  $\delta \in [0, 1/2]$ . As  $\delta$  varies, this model smoothly interpolates two well-studied models: the Gaussian Location Model for which  $\delta=0$  and the Gaussian Mixture Model for which  $\delta=1/2$ . Assuming that the estimator knows  $\delta$ , we establish a nearly minimax optimal (up to logarithmic factors) estimation error rate, as a function of  $\|\theta\|, \delta, d, n$ . We then provide an upper bound to the case of estimating  $\delta$ , assuming a (possibly inaccurate) knowledge of  $\theta$ . The bound is proved to be tight when  $\theta$  is an accurately known constant. These results are then combined to an algorithm which estimates  $\theta^*$  with  $\delta$  unknown a priori, and theoretical guarantees on its error are stated.

## [Learning Mixed Multinomial Logits with Provable Guarantees](#)

- Yiqun Hu Â· David Simchi-Levi Â· Zhenzhen Yan
- abstract@[open-review](#): A mixture of multinomial logits (MMNL) generalizes the single logit model, which is commonly used in predicting the probabilities of different outcomes. While extensive algorithms have been developed in the literature to learn MMNL models, theoretical results are limited. Built on the Frank-Wolfe (FW) method, we propose a new algorithm that learns both mixture weights and component-specific logit parameters with provable convergence guarantees for an arbitrary number of mixtures. Our algorithm utilizes historical choice data to generate a set of candidate choice probability vectors, each being close to the ground truth with a high probability. We further provide a sample complexity analysis to show that only a polynomial number of samples is required to secure the performance guarantee of our algorithm. Finally, we conduct simulation studies to evaluate the performance and demonstrate how to apply our algorithm to real-world applications.

## [Fairness without Demographics through Knowledge Distillation](#)

- Junyi Chai Â· Taeuk Jang Â· Xiaoqian Wang
- abstract@[open-review](#): Most of existing work on fairness assumes available demographic information in the training set. In practice, due to legal or privacy concerns, when demographic information is not available in the training set, it is crucial to find alternative objectives to ensure fairness. Existing work on fairness without demographics follows Rawlsian Max-Min fairness objective. However, such constraints could be too strict to achieve expected improvement in group fairness, and could lead to a great decrease in accuracy. In light of these limitations, in this paper, we propose to solve the problem from a new perspective, i.e., through knowledge distillation. Our method uses soft label from an overfitted teacher model as an alternative, and we show from preliminary experiments that soft labelling is beneficial for improving fairness. We analyze theoretically the fairness of our method, and we show that our method can be treated as an error-based reweighing. Experimental results on three datasets show that our method outperforms state-of-the-art alternatives, with notable improvements in group fairness and with relatively small decrease in accuracy.

## [Recurrent Memory Transformer](#)

- Aydar Bulatov Â· Yury Kuratov Â· Mikhail Burtsev

- abstract@[open-review](#): Transformer-based models show their effectiveness across multiple domains and tasks. The self-attention allows to combine information from all sequence elements into context-aware representations. However, global and local information has to be stored mostly in the same element-wise representations. Moreover, the length of an input sequence is limited by quadratic computational complexity of self-attention. In this work, we propose and study a memory-augmented segment-level recurrent Transformer (RMT). Memory allows to store and process local and global information as well as to pass information between segments of the long sequence with the help of recurrence. We implement a memory mechanism with no changes to Transformer model by adding special memory tokens to the input or output sequence. Then the model is trained to control both memory operations and sequence representations processing. Results of experiments show that RMT performs on par with the Transformer-XL on language modeling for smaller memory sizes and outperforms it for tasks that require longer sequence processing. We show that adding memory tokens to Tr-XL is able to improve its performance. This makes Recurrent Memory Transformer a promising architecture for applications that require learning of long-term dependencies and general purpose in memory processing, such as algorithmic tasks and reasoning.

## [Self-supervised surround-view depth estimation with volumetric feature fusion](#)

- Jung-Hee Kim · Junhwa Hur · Tien Phuoc Nguyen · Seong-Gyun Jeong
- abstract@[open-review](#): In this work, we propose a self-supervised depth estimation method using a unified volumetric feature encoded from surround-view. The proposed network architecture consists of three parts. First, given a set of surround-view images, the surround-view feature fusion module extracts image features from each view, aggregates the features into 3D space, and encodes them into a unified volumetric feature. We use a multilayer perceptron to refine the volumetric features, especially for features shared between multiple views. Second, we propose a depth fusion module that takes a specific viewpoint of interest from the volumetric feature and reconstructs a depth map at the corresponding view; therefore, our method is able to render scale-aware depth maps not only at known input camera views but also at any arbitrary rotated views. Lastly, assuming static camera extrinsics in the multi-camera system, we propose to estimate a single global motion according to a canonical camera coordinate system. The proposed method leverages 3D spatio-temporal context to learn metric-scale depth in a self-supervised manner. We adjust the intensity distribution around common image boundaries to avoid irregular photometric reconstruction errors. Through the extensive experiments on DDAD and nuScenes datasets, we show that our method outperforms the prior arts.

## [Algorithms with Prediction Portfolios](#)

- Michael Dinitz · Sungjin Im · Thomas Lavastida · Benjamin Moseley · Sergei Vassilvitskii
- abstract@[open-review](#): The research area of algorithms with predictions has seen recent success showing how to incorporate machine learning into algorithm design to improve performance when the predictions are correct, while retaining worst-case guarantees when they are not. Most previous work has assumed that the algorithm has access to a single predictor. However, in practice, there are many machine learning methods available, often with incomparable generalization guarantees, making it hard to pick a best method a priori. In this work we consider scenarios where multiple predictors are available to the algorithm and the question is how to best utilize them. Ideally, we would like the algorithm's performance to depend on the quality of the {em best} predictor. However, utilizing more predictions comes with a cost, since we now have to identify which prediction is best. We study the use of multiple predictors for a number of fundamental problems, including matching, load balancing, and non-clairvoyant scheduling, which have been well-studied in the single predictor setting. For each of these problems we introduce new algorithms that take advantage of multiple predictors, and prove bounds on the resulting performance.

## [Multi-fidelity Monte Carlo: a pseudo-marginal approach](#)

- Diana Cai · Ryan Adams
- abstract@[open-review](#): Markov Chain Monte Carlo (MCMC) is an established approach for uncertainty quantification and propagation in scientific applications. A key challenge in applying MCMC to scientific domains is computation: the target density of interest is often a function of expensive computations, such as a high-fidelity physical simulation, an intractable integral, or a slowly-converging iterative algorithm. Thus, using an MCMC algorithm with an expensive target density becomes impractical, as these expensive computations need to be evaluated at each iteration of the algorithm. In practice, these computations are often approximated via a cheaper, low-fidelity computation, leading to bias in the resulting target density. Multi-fidelity MCMC algorithms combine likelihoods of varying fidelities in order to obtain an approximate target density with lower computational cost. In this paper, we describe a class of asymptotically exact multi-fidelity MCMC algorithms for the setting where a sequence of likelihoods of increasing fidelity can be computed that approximates the high-fidelity likelihood. We take a pseudo-marginal MCMC approach for multi-fidelity inference that utilizes a randomized-fidelity unbiased estimator of the high-fidelity likelihood constructed via randomized truncation of a telescoping series of the low-fidelity sequence of models. Finally, we discuss and evaluate the proposed multi-fidelity MCMC approach on several applications, including log-Gaussian Cox process modeling, Bayesian ODE system identification, PDE-constrained optimization, and Gaussian process regression parameter inference.

## [Fine-tuning language models to find consensus among humans with diverse preferences](#)

- Michiel Bakker · Martin Chadwick · Hannah Sheahan · Michael Tessler · Lucy Campbell-Gillingham · Jan Balaguer · Nat McAleese · Amelia Glaese · John Aslanides · Matt Botvinick · Christopher Summerfield
- abstract@[open-review](#): Recent work in large language modeling (LLMs) has used fine-tuning to align outputs with the preferences of a prototypical user. This work assumes that human preferences are static and homogeneous across individuals, so that aligning to a single "generic" user will confer more general alignment. Here, we embrace the heterogeneity of human preferences to consider a different challenge: how might a machine help people with diverse views find agreement? We fine-tune a 70 billion parameter LLM to generate consensus statements that maximize the expected approval for a group of people with potentially diverse opinions. Human participants provide written opinions on thousands of questions touching on moral and political issues (e.g., "should we raise taxes on the rich?"), and rate the LLM's generated consensus statements for agreement and quality. A reward model is then trained to predict individual preferences, enabling it to quantify and rank consensus statements in terms of their appeal to the overall group, defined according to different aggregation (social welfare) functions. The model produces consensus statements that are preferred by human users over those from prompted LLMs ( $>70\%$ ) and significantly outperforms a tight fine-tuned baseline that lacks the final ranking step. Further, our best model consensus statements are preferred over the best human-generated opinions ( $>65\%$ ). We find that when we silently constructed consensus statements from only a subset of group members, those who were excluded were more likely to dissent, revealing the sensitivity of the consensus to individual contributions. These results highlight the potential to use LLMs to help groups of humans align their values with one another.

## [Robust Binary Models by Pruning Randomly-initialized Networks](#)

- Chen Liu · Ziqi Zhao · Sabine Sässstrunk · Mathieu Salzmann
- abstract@[open-review](#): Robustness to adversarial attacks was shown to require a larger model capacity, and thus a larger memory footprint. In this paper, we introduce an approach to obtain robust yet compact models by pruning randomly-initialized binary networks. Unlike adversarial training, which learns the model parameters, we initialize the model parameters as either  $+1\$$  or  $-1\$$ , keep them fixed, and find a subnetwork structure that is robust to attacks. Our method confirms the Strong Lottery Ticket Hypothesis in the presence of adversarial attacks, and extends this to binary networks. Furthermore, it yields more compact networks with competitive performance than existing works by 1) adaptively pruning the different network layers; 2) exploiting an effective binary initialization scheme; 3) incorporating a last batch normalization layer to improve training stability. Our experiments demonstrate that our approach not only always outperforms the state-of-the-art robust binary networks, but also can achieve accuracy better than full-precision ones in some datasets. Finally, we show the structured patterns of our pruned binary networks.

## Few-shot Relational Reasoning via Pretraining of Connection Subgraph Reconstruction

- Qian Huang · Hongyu Ren · Jure Leskovec
- abstract@[open-review](#): Few-shot knowledge graph (KG) completion task aims to perform inductive reasoning over the KG: given only a few support triplets of a new relation \$\bowtie\$ (e.g., (chop, \$\bowtie\$, kitchen), (read, \$\bowtie\$, library)), predict the query triplets of the same unseen relation \$\bowtie\$, e.g., (sleep, \$\bowtie\$, ?), with the existing knowledge in KG. Current approaches cast the problem in a meta-learning framework, where the model needs to be first jointly trained over many training few-shot tasks, each being defined by its own relation, so that learning/prediction on the target few-shot task can be effective. However, in real-world KGs, curating many training tasks is a challenging ad hoc process. Here we propose Connection Subgraph Reasoner (CSR), which can make predictions for the target few-shot task directly without the need for pre-training on the human curated set of training tasks. The key to CSR is that we explicitly model a shared connection subgraph between support and query triplets, as inspired by the principle of eliminative induction. To adapt to specific KG, we design a corresponding self-supervised pretraining scheme with the objective of reconstructing automatically sampled connection subgraphs. Our pretrained model can then be directly applied to target few-shot tasks without the need for training few-shot tasks. Extensive experiments on real KGs, including NELL, FB15K-237, and ConceptNet, demonstrate the effectiveness of our framework: we show that even a learning-free implementation of CSR can already perform competitively to existing methods on target few-shot tasks; with pretraining, CSR can achieve significant gains of up to 56% on the more challenging inductive few-shot tasks where the entities are also unseen during (pre)training.

## Transformer Memory as a Differentiable Search Index

- Yi Tay · Vinh Tran · Mostafa Dehghani · Jianmo Ni · Dara Bahri · Harsh Mehta · Zhen Qin · Kai Hui · Zhe Zhao · Jai Gupta · Tal Schuster · William Cohen · Donald Metzler
- abstract@[open-review](#): In this paper, we demonstrate that information retrieval can be accomplished with a single Transformer, in which all information about the corpus is encoded in the parameters of the model. To this end, we introduce the Differentiable Search Index (DSI), a new paradigm that learns a text-to-text model that maps string queries directly to relevant docids; in other words, a DSI model answers queries directly using only its parameters, dramatically simplifying the whole retrieval process. We study variations in how documents and their identifiers are represented, variations in training procedures, and the interplay between models and corpus sizes. Experiments demonstrate that given appropriate design choices, DSI significantly outperforms strong baselines such as dual encoder models. Moreover, DSI demonstrates strong generalization capabilities, outperforming a BM25 baseline in a zero-shot setup.

## Task Discovery: Finding the Tasks that Neural Networks Generalize on

- Andrei Atanov · Andrey Filatov · Teresa Yeo · Ajay Sohmshetty · Amir Zamir
- abstract@[open-review](#): When developing deep learning models, we usually decide what task we want to solve then search in the space of models in order to design one that generalizes well on this task. An intriguing question would be: what if, instead of fixing the task and searching in the model space, we fix the model and search in the task space? Can we find tasks that the model generalizes on? What do they look like, or do they show anything? This is the question we address in this paper. We propose a task discovery framework that automatically finds examples of such tasks via optimizing a generalization-based quantity called agreement score. With this framework, we demonstrate that the same set of images can allow for many tasks on which neural networks generalize well. The understandings from task discovery can also provide a tool to shed more light on deep learning and its failure modes: as an example, we show that the discovered tasks can be used to generate ``adversarial train-test splits'' which make a model fail at test time, without changing the pixels or labels, but only by selecting how the datapoints should be split between training and testing.

## Identification, Amplification and Measurement: A bridge to Gaussian Differential Privacy

- Yi Liu · Ke Sun · Bei Jiang · Linglong Kong
- abstract@[open-review](#): Gaussian differential privacy (GDP) is a single-parameter family of privacy notions that provides coherent guarantees to avoid the exposure of sensitive individual information. Despite the extra interpretability and tighter bounds under composition GDP provides, many widely used mechanisms (e.g., the Laplace mechanism) inherently provide GDP guarantees but often fail to take advantage of this new framework because their privacy guarantees were derived under a different background. In this paper, we study the asymptotic properties of privacy profiles and develop a simple criterion to identify algorithms with GDP properties. We propose an efficient method for GDP algorithms to narrow down possible values of an optimal privacy measurement,  $\mu$  with an arbitrarily small and quantifiable margin of error. For non GDP algorithms, we provide a post-processing procedure that can amplify existing privacy guarantees to meet the GDP condition. As applications, we compare two single-parameter families of privacy notions,  $\epsilon$ -DP, and  $\mu$ -GDP, and show that all  $\epsilon$ -DP algorithms are intrinsically also GDP. Lastly, we show that the combination of our measurement process and the composition theorem of GDP is a powerful and convenient tool to handle compositions compared to the traditional standard and advanced composition theorems.

## Learning Optimal Flows for Non-Equilibrium Importance Sampling

- Yu Cao · Eric Vanden-Eijnden
- abstract@[open-review](#): Many applications in computational sciences and statistical inference require the computation of expectations with respect to complex high-dimensional distributions with unknown normalization constants, as well as the estimation of these constants. Here we develop a method to perform these calculations based on generating samples from a simple base distribution, transporting them along the flow generated by a velocity field, and performing averages along these flowlines. This non-equilibrium importance sampling (NEIS) strategy is straightforward to implement, and can be used for calculations with arbitrary target distributions. On the theory side we discuss how to tailor the velocity field to the target and establish general conditions under which the proposed estimator is a perfect estimator, with zero-variance. We also draw connections between NEIS and approaches based on mapping a base distribution onto a target via a transport map. On the computational side we show how to use deep learning to represent the velocity field by a neural network and train it towards the zero variance optimum. These results are illustrated numerically on benchmark examples (with dimension up to 10), where we show that training the velocity field can decrease the variance of the NEIS estimator by up to 6 orders of magnitude compared to a vanilla estimator. We also show that NEIS with optimized flows achieves significant variance reduction on these examples than Nealâ€™s annealed importance sampling (AIS).

## Incentivizing Combinatorial Bandit Exploration

- Xinyan Hu · Dung Ngo · Aleksandrs Slivkins · Steven Wu
- abstract@[open-review](#): Consider a bandit algorithm that recommends actions to self-interested users in a recommendation system. The users are free to choose other actions and need to be incentivized to follow the algorithm's recommendations. While the users prefer to exploit, the algorithm can incentivize them to explore by leveraging the information collected from the previous users. All published work on this problem, known as incentivized exploration, focuses on small, unstructured action sets and mainly targets the case when the users' beliefs are independent across actions. However, realistic exploration problems often feature large, structured action sets and highly correlated beliefs. We focus on a paradigmatic exploration problem with structure: combinatorial semi-bandits. We prove that Thompson Sampling, when applied to combinatorial semi-bandits, is incentive-compatible when initialized with a sufficient number of samples of each arm (where this number is determined in advance by the Bayesian prior). Moreover, we design incentive-compatible algorithms for collecting the initial samples.

## Instance-optimal PAC Algorithms for Contextual Bandits

- Zhaoqi Li · Lillian Ratliff · houssam nassif · Kevin Jamieson · Lalit Jain
- abstract@[open-review](#): In the stochastic contextual bandit setting, regret-minimizing algorithms have been extensively researched, but their instance-minimizing best-arm identification counterparts remain seldom studied. In this work, we focus on the stochastic bandit problem in the  $(\epsilon, \delta)$ -PAC setting: given a policy class  $\Pi$  the goal of the learner is to return a policy  $\pi_i \in \Pi$  whose expected reward is within  $\epsilon$  of the optimal policy with probability greater than  $1 - \delta$ . We characterize the first instance-dependent PAC sample complexity of contextual bandits through a quantity  $\rho_{\Pi}$ , and provide matching upper and lower bounds in terms of  $\rho_{\Pi}$  for the agnostic and linear contextual best-arm identification settings. We show that no algorithm can be simultaneously minimax-optimal for regret minimization and instance-dependent PAC for best-arm identification. Our main result is a new instance-optimal and computationally efficient algorithm that relies on a polynomial number of calls to a cost-sensitive classification oracle.

## [Object Representations as Fixed Points: Training Iterative Refinement Algorithms with Implicit Differentiation](#)

- Michael Chang · Tom Griffiths · Sergey Levine
- abstract@[open-review](#): Current work in object-centric learning has been motivated by developing learning algorithms that infer independent and symmetric entities from the perceptual input. This often requires the use iterative refinement procedures that break symmetries among equally plausible explanations for the data, but most prior works differentiate through the unrolled refinement process, which can make optimization exceptionally challenging. In this work, we observe that such iterative refinement methods can be made differentiable by means of the implicit function theorem, and develop an implicit differentiation approach that improves the stability and tractability of training such models by decoupling the forward and backward passes. This connection enables us to apply recent advances in optimizing implicit layers to not only improve the stability and optimization of the slot attention module in SLATE, a state-of-the-art method for learning entity representations, but do so with constant space and time complexity in backpropagation and only one additional line of code.

## [Efficient Multi-agent Communication via Self-supervised Information Aggregation](#)

- Cong Guan · Feng Chen · Lei Yuan · Chenghe Wang · Hao Yin · Zongzhang Zhang · Yang Yu
- abstract@[open-review](#): Efficiently utilizing messages from others can improve coordination in cooperative Multi-agent Reinforcement Learning (MARL). Previous works mainly focus on generating meaningful messages or selecting the most relevant message for decision-making. However, they simply combine the received messages with local information and train the policy in an end-to-end way, neglecting the aggregation of multiple messages. We argue that an agent can coordinate better if it can learn to aggregate the received message efficiently. To that end, we propose  $M$ -agent communication via  $S$ -elf-supervised  $I$ nformation  $A$ ggregation (MASIA), with which agents can ground the received message into compact representations and extract the most relevant part to augment the local policy. Specifically, a permutation invariant message encoder is first applied to compact the raw message. We then optimize it by reconstructing the true state along with predicting its latent state representations multiple steps into the future in a self-supervised manner. A message extraction mechanism is finally employed to obtain the most meaningful message for decision-making. Sufficient empirical results demonstrate that our method is agnostic to specific MARL algorithms, and significantly outperforms strong baselines and achieves excellent performance on multiple cooperative MARL tasks for various task settings.

## [Evaluated CMI Bounds for Meta Learning: Tightness and Expressiveness](#)

- Fredrik Hellström · Giuseppe Durisi
- abstract@[open-review](#): Recent work has established that the conditional mutual information (CMI) framework of Steinke and Zakynthinou (2020) is expressive enough to capture generalization guarantees in terms of algorithmic stability, VC dimension, and related complexity measures for conventional learning (Harutyunyan et al., 2021, Haghifam et al., 2021). Hence, it provides a unified method for establishing generalization bounds. In meta learning, there has so far been a divide between information-theoretic results and results from classical learning theory. In this work, we take a first step toward bridging this divide. Specifically, we present novel generalization bounds for meta learning in terms of the evaluated CMI (e-CMI). To demonstrate the expressiveness of the e-CMI framework, we apply our bounds to a representation learning setting, with  $n$  samples from  $\hat{n}$  tasks parameterized by functions of the form  $f_i \circ h$ . Here, each  $f_i(\cdot)$  is a task-specific function, and  $h(\cdot)$  is the shared representation. For this setup, we show that the e-CMI framework yields a bound that scales as  $\sqrt{\mathcal{C}(\mathcal{H})/\hat{n}} + \mathcal{C}(\mathcal{F})/n$ , where  $\mathcal{C}(\cdot)$  denotes a complexity measure of the hypothesis class. This scaling behavior coincides with the one reported in Tripuraneni et al. (2020) using Gaussian complexity.

## [Learning to Reconstruct Missing Data from Spatiotemporal Graphs with Sparse Observations](#)

- Ivan Marisca · Andrea Cini · Cesare Alippi
- abstract@[open-review](#): Modeling multivariate time series as temporal signals over a (possibly dynamic) graph is an effective representational framework that allows for developing models for time series analysis. In fact, discrete sequences of graphs can be processed by autoregressive graph neural networks to recursively learn representations at each discrete point in time and space. Spatiotemporal graphs are often highly sparse, with time series characterized by multiple, concurrent, and even long sequences of missing data, e.g., due to the unreliable underlying sensor network. In this context, autoregressive models can be brittle and exhibit unstable learning dynamics. The objective of this paper is, then, to tackle the problem of learning effective models to reconstruct, i.e., impute, missing data points by conditioning the reconstruction only on the available observations. In particular, we propose a novel class of attention-based architectures that, given a set of highly sparse discrete observations, learn a representation for points in time and space by exploiting a spatiotemporal diffusion architecture aligned with the imputation task. Representations are trained end-to-end to reconstruct observations w.r.t. the corresponding sensor and its neighboring nodes. Compared to the state of the art, our model handles sparse data without propagating prediction errors or requiring a bidirectional model to encode forward and backward time dependencies. Empirical results on representative benchmarks show the effectiveness of the proposed method.

## [Lossless Compression of Deep Neural Networks: A High-dimensional Neural Tangent Kernel Approach](#)

- lingyu gu · Yongqi Du · yuan zhang · Di Xie · Shiliang Pu · Robert Qiu · Zhenyu Liao
- abstract@[open-review](#): Modern deep neural networks (DNNs) are extremely powerful; however, this comes at the price of increased depth and having more parameters per layer, making their training and inference more computationally challenging. In an attempt to address this key limitation, efforts have been devoted to the compression (e.g., sparsification and/or quantization) of these large-scale machine learning models, so that they can be deployed on low-power IoT devices. In this paper, building upon recent research advances in the neural tangent kernel (NTK) and random matrix theory, we provide a novel compression approach to wide and fully-connected deep neural nets. Specifically, we demonstrate that in the high-dimensional regime where the number of data points  $n$  and their dimension  $p$  are both large, and under a Gaussian mixture model for the data, there exists asymptotic equivalence between the NTK matrices for a large family of DNN models. This theoretical result enables "lossless" compression of a given DNN to be performed, in the sense that the compressed network yields asymptotically the same NTK as the original (dense and unquantized) network, with its weights and activations taking values only in  $[0, 1]$  up to scaling. Experiments on both synthetic and real-world data are conducted to support the numerical advantages of the proposed method.

## [Sharing Knowledge for Meta-learning with Feature Descriptions](#)

- Tomoharu Iwata · Atsutoshi Kumagai

- abstract@[open-review](#): Language is an important tool for humans to share knowledge. We propose a meta-learning method that shares knowledge across supervised learning tasks using feature descriptions written in natural language, which have not been used in the existing meta-learning methods. The proposed method improves the predictive performance on unseen tasks with a limited number of labeled data by meta-learning from various tasks. With the feature descriptions, we can find relationships across tasks even when their feature spaces are different. The feature descriptions are encoded using a language model pretrained with a large corpus, which enables us to incorporate human knowledge stored in the corpus into meta-learning. In our experiments, we demonstrate that the proposed method achieves better predictive performance than the existing meta-learning methods using a wide variety of real-world datasets provided by the statistical office of the EU and Japan.

## [Understanding Deep Contrastive Learning via Coordinate-wise Optimization](#)

- Yuandong Tian
- abstract@[open-review](#): We show that Contrastive Learning (CL) under a broad family of loss functions (including InfoNCE) has a unified formulation of coordinate-wise optimization on the network parameter  $\boldsymbol{\theta}$  and pairwise importance  $\alpha$ , where the max player  $\boldsymbol{\theta}$  learns representation for contrastiveness, and the min player  $\alpha$  puts more weights on pairs of distinct samples that share similar representations. The resulting formulation, called  $\alpha$ CL, unifies not only various existing contrastive losses, which differ by how sample-pair importance  $\alpha$  is constructed, but also is able to extrapolate to give novel contrastive losses beyond popular ones, opening a new avenue of contrastive loss design. These novel losses yield comparable (or better) performance on CIFAR10 and STL-10 than classic InfoNCE. Furthermore, we also analyze the max player in detail: we prove that with fixed  $\alpha$ , max player is equivalent to Principal Component Analysis (PCA) for deep linear network, and almost all local minima are global and rank-1, recovering optimal PCA solutions. Finally, we extend our analysis on max player to 2-layer ReLU networks, showing that its fixed points can have higher ranks.

## [Simple Unsupervised Object-Centric Learning for Complex and Naturalistic Videos](#)

- Gautam Singh · Yi-Fu Wu · Sungjin Ahn
- abstract@[open-review](#): Unsupervised object-centric learning aims to represent the modular, compositional, and causal structure of a scene as a set of object representations and thereby promises to resolve many critical limitations of traditional single-vector representations such as poor systematic generalization. Although there have been many remarkable advances in recent years, one of the most critical problems in this direction has been that previous methods work only with simple and synthetic scenes but not with complex and naturalistic images or videos. In this paper, we propose STEVE, an unsupervised model for object-centric learning in videos. Our proposed model makes a significant advancement by demonstrating its effectiveness on various complex and naturalistic videos unprecedented in this line of research. Interestingly, this is achieved by neither adding complexity to the model architecture nor introducing a new objective or weak supervision. Rather, it is achieved by a surprisingly simple architecture that uses a transformer-based image decoder conditioned on slots and the learning objective is simply to reconstruct the observation. Our experiment results on various complex and naturalistic videos show significant improvements compared to the previous state-of-the-art.

## [Active Learning with Safety Constraints](#)

- Romain Camilleri · Kevin Jamieson · Jamie Morgenstern · Lalit Jain · Andrew Wagenmaker
- abstract@[open-review](#): Active learning methods have shown great promise in reducing the number of samples necessary for learning. As automated learning systems are adopted into real-time, real-world decision-making pipelines, it is increasingly important that such algorithms are designed with safety in mind. In this work we investigate the complexity of learning the best safe decision in interactive environments. We reduce this problem to a safe linear bandits problem, where our goal is to find the best arm satisfying certain (unknown) safety constraints. We propose an adaptive experimental design-based algorithm, which we show efficiently trades off between the difficulty of showing an arm is unsafe vs suboptimal. To our knowledge, our results are the first on best-arm identification in linear bandits with safety constraints. In practice, we demonstrate that this approach performs well on synthetic and real world datasets.

## [Constrained GPI for Zero-Shot Transfer in Reinforcement Learning](#)

- Jaekyeom Kim · Seohong Park · Gunhee Kim
- abstract@[open-review](#): For zero-shot transfer in reinforcement learning where the reward function varies between different tasks, the successor features framework has been one of the popular approaches. However, in this framework, the transfer to new target tasks with generalized policy improvement (GPI) relies on only the source successor features [3] or additional successor features obtained solely with the function approximators' generalization to novel inputs [8]. The goal of this work is to improve the transfer by further bounding the value approximation errors of successor features on the new target tasks. To this end, we first present lower and upper bounds of optimal values for the novel tasks that are expressible as linear combinations of source tasks given the reward decomposition structure. Based on the bounds, we then propose constrained GPI as a simple test-time approach that can improve transfer by constraining action-value approximation errors on new target tasks. With experiments in the Scavenger and Reacher environments, we show that the proposed constrained GPI significantly outperforms the prior GPI's transfer performance.

## [A Unified Hard-Constraint Framework for Solving Geometrically Complex PDEs](#)

- Songming Liu · Hao Zhongkai · Chengyang Ying · Hang Su · Jun Zhu · Ze Cheng
- abstract@[open-review](#): We present a unified hard-constraint framework for solving geometrically complex PDEs with neural networks, where the most commonly used Dirichlet, Neumann, and Robin boundary conditions (BCs) are considered. Specifically, we first introduce the extra fields" from the mixed finite element method to reformulate the PDEs so as to equivalently transform the three types of BCs into linear forms. Based on the reformulation, we derive the general solutions of the BCs analytically, which are employed to construct an ansatz that automatically satisfies the BCs. With such a framework, we can train the neural networks without adding extra loss terms and thus efficiently handle geometrically complex PDEs, alleviating the unbalanced competition between the loss terms corresponding to the BCs and PDEs. We theoretically demonstrate that the extra fields" can stabilize the training process. Experimental results on real-world geometrically complex PDEs showcase the effectiveness of our method compared with state-of-the-art baselines.

## [Rethinking Knowledge Graph Evaluation Under the Open-World Assumption](#)

- Haotong Yang · Zhouchen Lin · Muhan Zhang
- abstract@[open-review](#): Most knowledge graphs (KGs) are incomplete, which motivates one important research topic on automatically complementing knowledge graphs. However, evaluation of knowledge graph completion (KGC) models often ignores the incompleteness---facts in the test set are ranked against all unknown triplets which may contain a large number of missing facts not included in the KG yet. Treating all unknown triplets as false is called the closed-world assumption. This closed-world assumption might negatively affect the fairness and consistency of the evaluation metrics. In this paper, we study KGC evaluation under a more realistic setting, namely the open-world assumption, where unknown triplets are considered to include many missing facts not included in the training or test sets. For the currently most used metrics such as mean reciprocal rank (MRR) and Hits@K, we point out that their behavior may be unexpected under the open-world assumption. Specifically, with not many missing facts, their numbers show a logarithmic trend with respect to the true strength of the model, and thus, the metric increase could be insignificant in terms of reflecting the true model improvement. Further, considering the variance, we show that the degradation in the reported numbers may result in incorrect comparisons between different models,

where stronger models may have lower metric numbers. We validate the phenomenon both theoretically and experimentally. Finally, we suggest possible causes and solutions for this problem.

## [So3krates - Self-attention for higher-order geometric interactions on arbitrary length-scales](#)

- Thorben Frank · Oliver Unke · Klaus-Robert Müller
- abstract@[open-review](#): The application of machine learning methods in quantum chemistry has enabled the study of numerous chemical phenomena, which are computationally intractable with traditional ab-initio methods. However, some quantum mechanical properties of molecules and materials depend on non-local electronic effects, which are often neglected due to the difficulty of modeling them efficiently. This work proposes a modified attention mechanism adapted to the underlying physics, which allows to recover the relevant non-local effects. Namely, we introduce spherical harmonic coordinates (SPHCs) to reflect higher-order geometric information for each atom in a molecule, enabling a non-local formulation of attention in the SPHC space. Our proposed model So3krates -- a self-attention based message passing neural network -- uncouples geometric information from atomic features, making them independently amenable to attention mechanisms. We show that in contrast to other published methods, So3krates is able to describe non-local quantum mechanical effects over arbitrary length scales. Further, we find evidence that the inclusion of higher-order geometric correlations increases data efficiency and improves generalization. So3krates matches or exceeds state-of-the-art performance on popular benchmarks, notably, requiring a significantly lower number of parameters (0.25--0.4x) while at the same time giving a substantial speedup (6--14x for training and 2--11x for inference) compared to other models.

## [Group Meritocratic Fairness in Linear Contextual Bandits](#)

- Riccardo Grazzi · Arya Akhavan · Massimiliano Pontil · John IF Falk · Leonardo Cellia
- abstract@[open-review](#): We study the linear contextual bandit problem where an agent has to select one candidate from a pool and each candidate belongs to a sensitive group. In this setting, candidates' rewards may not be directly comparable between groups, for example when the agent is an employer hiring candidates from different ethnic groups and some groups have a lower reward due to discriminatory bias and/or social injustice. We propose a notion of fairness that states that the agent's policy is fair when it selects a candidate with highest relative rank, which measures how good the reward is when compared to candidates from the same group. This is a very strong notion of fairness, since the relative rank is not directly observed by the agent and depends on the underlying reward model and on the distribution of rewards. Thus we study the problem of learning a policy which approximates a fair policy under the condition that the contexts are independent between groups and the distribution of rewards of each group is absolutely continuous. In particular, we design a greedy policy which at each round constructs a ridge regression estimator from the observed context-reward pairs, and then computes an estimate of the relative rank of each candidate using the empirical cumulative distribution function. We prove that the greedy policy achieves, after  $T$  rounds, up to log factors and with high probability, a fair pseudo-regret of order  $\sqrt{dT}$ , where  $d$  is the dimension of the context vectors. The policy also satisfies demographic parity at each round when averaged over all possible information available before the selection. We finally show with a proof of concept simulation that our policy achieves sub-linear fair pseudo-regret also in practice.

## [Beyond IID: data-driven decision-making in heterogeneous environments](#)

- Omar Besbes · Will Ma · Omar Mouchtaki
- abstract@[open-review](#): In this work, we study data-driven decision-making and depart from the classical identically and independently distributed (i.i.d.) assumption. We present a new framework in which historical samples are generated from unknown and different distributions, which we dub \textit{heterogeneous environments}. These distributions are assumed to lie in a heterogeneity ball with known radius and centered around the (also) unknown future (out-of-sample) distribution on which the performance of a decision will be evaluated. We quantify the asymptotic worst-case regret that is achievable by central data-driven policies such as Sample Average Approximation, but also by rate-optimal ones, as a function of the radius of the heterogeneity ball. Our work shows that the type of achievable performance varies considerably across different combinations of problem classes and notions of heterogeneity. We demonstrate the versatility of our framework by comparing achievable guarantees for the heterogeneous version of widely studied data-driven problems such as pricing, ski-rental, and news-vendor. En route, we establish a new connection between data-driven decision-making and distributionally robust optimization.

## [Generalization Bounds for Gradient Methods via Discrete and Continuous Prior](#)

- Jian Li · Xuanyuan Luo
- abstract@[open-review](#): Proving algorithm-dependent generalization error bounds for gradient-type optimization methods has attracted significant attention recently in learning theory. However, most existing trajectory-based analyses require either restrictive assumptions on the learning rate (e.g., fast decreasing learning rate), or continuous injected noise (such as the Gaussian noise in Langevin dynamics). In this paper, we introduce a new discrete data-dependent prior to the PAC-Bayesian framework, and prove a high probability generalization bound of order  $O(\frac{1}{n}) \cdot \sum_{t=1}^T (\gamma_t / \Delta_{\epsilon_t})^2 \left( \mathbb{E}[\|g_t\|^2] - \mathbb{E}[\|g_t\|]^2 \right)$  for Floored GD (i.e. a version of gradient descent with precision level  $\Delta_{\epsilon_t}$ ), where  $n$  is the number of training samples,  $\gamma_t$  is the learning rate at step  $t$ ,  $\|g_t\|$  is roughly the difference of the gradient computed using all samples and that using only prior samples.  $\mathbb{E}[\|g_t\|]$  is upper bounded by  $L$  and typical much smaller than the gradient norm  $\|\nabla f(W_t)\|$ . We remark that our bound holds for nonconvex and nonsmooth scenarios. Moreover, our theoretical results provide numerically favorable upper bounds of testing errors (e.g.,  $0.037$  on MNIST). Using similar technique, we can also obtain new generalization bounds for a certain variant of SGD. Furthermore, we study the generalization bounds for gradient Langevin Dynamics (GLD). Using the same framework with a carefully constructed continuous prior, we show a new high probability generalization bound of order  $O(\frac{1}{n} + \frac{L^2}{n^2} \sum_{t=1}^T (\gamma_t / \sigma_t)^2)$  for GLD. The new  $1/n^2$  rate is due to the concentration of the difference between the gradient of training samples and that of the prior.

## [Log-Concave and Multivariate Canonical Noise Distributions for Differential Privacy](#)

- Jordan Awan · Jinshuo Dong
- abstract@[open-review](#): A canonical noise distribution (CNDs) is an additive mechanism designed to satisfy  $\epsilon$ -differential privacy ( $\epsilon$ -DP), without any wasted privacy budget.  $\epsilon$ -DP is a hypothesis testing-based formulation of privacy phrased in terms of \textit{tradeoff functions}, which captures the difficulty of a hypothesis test. In this paper, we consider the existence and construction of log-concave CNDs as well as multivariate CNDs. Log-concave distributions are important to ensure that higher outputs of the mechanism correspond to higher input values, whereas multivariate noise distributions are important to ensure that a joint release of multiple outputs has a tight privacy characterization. We show that the existence and construction of CNDs for both types of problems is related to whether the tradeoff function can be decomposed by functional composition (related to group privacy) or mechanism composition. In particular, we show that pure  $\epsilon$ -DP cannot be decomposed in either way and that there is neither a log-concave CND nor any multivariate CND for  $\epsilon$ -DP. On the other hand, we show that Gaussian-DP,  $(0, \delta)$ -DP, and Laplace-DP each have both log-concave and multivariate CNDs.

## [Benign Overfitting in Two-layer Convolutional Neural Networks](#)

- Yuan Cao · Zixiang Chen · Misha Belkin · Quanquan Gu
- abstract@[open-review](#): Modern neural networks often have great expressive power and can be trained to overfit the training data, while still achieving a good test performance. This phenomenon is referred to as "benign overfitting". Recently, there emerges a line of works studying "benign

overfitting from the theoretical perspective. However, they are limited to linear models or kernel/random feature models, and there is still a lack of theoretical understanding about when and how benign overfitting occurs in neural networks. In this paper, we study the benign overfitting phenomenon in training a two-layer convolutional neural network (CNN). We show that when the signal-to-noise ratio satisfies a certain condition, a two-layer CNN trained by gradient descent can achieve arbitrarily small training and test loss. On the other hand, when this condition does not hold, overfitting becomes harmful and the obtained CNN can only achieve a constant level test loss. These together demonstrate a sharp phase transition between benign overfitting and harmful overfitting, driven by the signal-to-noise ratio. To the best of our knowledge, this is the first work that precisely characterizes the conditions under which benign overfitting can occur in training convolutional neural networks.

## [Using Mixup as a Regularizer Can Surprisingly Improve Accuracy and Out Distribution Robustness](#)

- Francesco Pinto · Harry Yang · Ser Nam Lim · Philip Torr · Puneet Dokania
- abstract@[open-review](#): We show that the effectiveness of the well celebrated Mixup~\citet{zhang2018mixup} can be further improved if instead of using it as the sole learning objective, it is being utilized as an additional regularizer to the standard cross-entropy loss. This simple change not just provides much improved accuracy but also significantly improves the quality of the predictive uncertainty estimation of Mixup in most cases under various forms of covariate shifts and out-of-distribution detection experiments. Note, standard Mixup otherwise yield much degraded performance on out-of-distribution detection experiments, perhaps, as we show empirically, because of its tendency to learn models that throughout exhibit high-entropy which makes it difficult to differentiate between in-distribution and out-distribution samples. To show the efficacy of RegMixup (our approach), we provide thorough analysis and experiments on vision datasets (CIFAR-10/100 and ImageNet) and compare it with a suite of well-known approaches for reliable uncertainty estimation.

## [The First Optimal Algorithm for Smooth and Strongly-Convex-Strongly-Concave Minimax Optimization](#)

- Dmitry Kovalev · Alexander Gasnikov
- abstract@[open-review](#): In this paper, we revisit the smooth and strongly-convex-strongly-concave minimax optimization problem. Zhang et al. (2021) and Ibrahim et al. (2020) established the lower bound  $\Omega(\sqrt{\kappa_x \kappa_y} \log \frac{1}{\epsilon})$  on the number of gradient evaluations required to find an  $\mu$ -accurate solution, where  $\kappa_x$  and  $\kappa_y$  are condition numbers for the strong convexity and strong concavity assumptions. However, the existing state-of-the-art methods do not match this lower bound: algorithms of Lin et al. (2020) and Wang and Li (2020) have gradient evaluation complexity  $\mathcal{O}(\sqrt{\kappa_x \kappa_y} \log^3 \frac{1}{\epsilon})$  and  $\mathcal{O}(\sqrt{\kappa_x \kappa_y} \log^3 (\kappa_x \kappa_y) \log \frac{1}{\epsilon})$ , respectively. We fix this fundamental issue by providing the first algorithm with  $\mathcal{O}(\sqrt{\kappa_x \kappa_y} \log \frac{1}{\epsilon})$  gradient evaluation complexity. We design our algorithm in three steps: (i) we reformulate the original problem as a minimization problem via the pointwise conjugate function; (ii) we apply a specific variant of the proximal point algorithm to the reformulated problem; (iii) we compute the proximal operator inexactly using the optimal algorithm for operator norm reduction in monotone inclusions.

## [Hiding Images in Deep Probabilistic Models](#)

- Haoyu Chen · Linqi Song · Zhenxing Qian · Xinpeng Zhang · Kede Ma
- abstract@[open-review](#): Data hiding with deep neural networks (DNNs) has experienced impressive successes in recent years. A prevailing scheme is to train an autoencoder, consisting of an encoding network to embed (or transform) secret messages in (or into) a carrier, and a decoding network to extract the hidden messages. This scheme may suffer from several limitations regarding practicability, security, and embedding capacity. In this work, we describe a different computational framework to hide images in deep probabilistic models. Specifically, we use a DNN to model the probability density of cover images, and hide a secret image in one particular location of the learned distribution. As an instantiation, we adopt a SinGAN, a pyramid of generative adversarial networks (GANs), to learn the patch distribution of one cover image. We hide the secret image by fitting a deterministic mapping from a fixed set of noise maps (generated by an embedding key) to the secret image during patch distribution learning. The stego SinGAN, behaving as the original SinGAN, is publicly communicated; only the receiver with the embedding key is able to extract the secret image. We demonstrate the feasibility of our SinGAN approach in terms of extraction accuracy and model security. Moreover, we show the flexibility of the proposed method in terms of hiding multiple images for different receivers and obfuscating the secret image.

## [Injecting Domain Knowledge from Empirical Interatomic Potentials to Neural Networks for Predicting Material Properties](#)

- Zeren Shui · Daniel Karls · Mingjian Wen · ilia Nikiforov · Ellad Tadmor · George Karypis
- abstract@[open-review](#): For decades, atomistic modeling has played a crucial role in predicting the behavior of materials in numerous fields ranging from microprocessor design to drug discovery. The most accurate methods in this domain are rooted in first-principles quantum mechanical calculations such as density functional theory (DFT). Because these methods have remained computationally prohibitive, practitioners have traditionally focused on defining physically motivated closed-form expressions known as empirical interatomic potentials (EIPs) that approximately model the interactions between atoms in materials. In recent years, neural network (NN)-based potentials trained on quantum mechanical (DFT-labeled) data have emerged as a more accurate alternative to conventional EIPs. However, the generalizability of these models relies heavily on the amount of labeled training data, which is often still insufficient to generate models suitable for general-purpose applications. In this paper, we propose two generic strategies that take advantage of unlabeled training instances to inject domain knowledge from conventional EIPs to NNs in order to increase their generalizability. The first strategy, based on weakly supervised learning, trains an auxiliary classifier on EIPs and selects the best-performing EIP to generate energies to supplement the ground-truth DFT energies in training the NN. The second strategy, based on transfer learning, first pretrains the NN on a large set of easily obtainable EIP energies, and then fine-tunes it on ground-truth DFT energies. Experimental results on three benchmark datasets demonstrate that the first strategy improves baseline NN performance by 5% to 51% while the second improves baseline performance by up to 55%. Combining them further boosts performance.

## [Capturing Failures of Large Language Models via Human Cognitive Biases](#)

- Erik Jones · Jacob Steinhardt
- abstract@[open-review](#): Large language models generate complex, open-ended outputs: instead of outputting a class label they write summaries, generate dialogue, or produce working code. In order to assess the reliability of these open-ended generation systems, we aim to identify qualitative categories of erroneous behavior, beyond identifying individual errors. To hypothesize and test for such qualitative errors, we draw inspiration from human cognitive biases---systematic patterns of deviation from rational judgement. Specifically, we use cognitive biases as motivation to (i) generate hypotheses for problems that models may have, and (ii) develop experiments that elicit these problems. Using code generation as a case study, we find that OpenAI's Codex errs predictably based on how the input prompt is framed, adjusts outputs towards anchors, and is biased towards outputs that mimic frequent training examples. We then use our framework to elicit high-impact errors such as incorrectly deleting files. Our results indicate that experimental methodology from cognitive science can help characterize how machine learning systems behave.

## [Symmetry-induced Disentanglement on Graphs](#)

- Giangiacomo Mercatali · Andre Freitas · Vikas Garg
- abstract@[open-review](#): Learning disentangled representations is important for unraveling the underlying complex interactions between latent generative factors. Disentanglement has been formalized using a symmetry-centric notion for unstructured spaces, however, graphs have eluded a similarly rigorous treatment. We fill this gap with a new notion of conditional symmetry based disentanglement, and leverage tools from Lie algebras to encode graph

properties into subgroups using suitable adaptations of generative models such as Variational Autoencoders. Unlike existing works on disentanglement, proposed models can learn to segregate the latent space into uncoupled and entangled parts. Experiments on synthetic and real datasets reveal the ability of these models to learn effective disengaged representations, and improve performance on downstream tasks such as few-shot classification and molecular generation.

## [Provably Feedback-Efficient Reinforcement Learning via Active Reward Learning](#)

- Dingwen Kong · Lin Yang
- abstract@[open-review](#): An appropriate reward function is of paramount importance in specifying a task in reinforcement learning (RL). Yet, it is known to be extremely challenging in practice to design a correct reward function for even simple tasks. Human-in-the-loop (HiL) RL allows humans to communicate complex goals to the RL agent by providing various types of feedback. However, despite achieving great empirical successes, HiL RL usually requires \emph{too much} feedback from a human teacher and also suffers from insufficient theoretical understanding. In this paper, we focus on addressing this issue from a theoretical perspective, aiming to provide provably feedback-efficient algorithmic frameworks that take human-in-the-loop to specify rewards of given tasks. We provide an \emph{active-learning}-based RL algorithm that first explores the environment without specifying a reward function and then asks a human teacher for only a few queries about the rewards of a task at some state-action pairs. After that, the algorithm guarantees to provide a nearly optimal policy for the task with high probability. We show that, even with the presence of random noise in the feedback, the algorithm only takes  $\tilde{O}(H\dim_{\mathcal{R}}^2)$  queries on the reward function to provide an  $\epsilon$ -optimal policy for any  $\epsilon > 0$ . Here  $H$  is the horizon of the RL environment, and  $\dim_{\mathcal{R}}$  specifies the complexity of the function class representing the reward function. In contrast, standard RL algorithms require to query the reward function for at least  $\Omega(\operatorname{poly}(d, 1/\epsilon))$  state-action pairs where  $d$  depends on the complexity of the environmental transition.

## [Nearly Optimal Best-of-Both-Worlds Algorithms for Online Learning with Feedback Graphs](#)

- Shinji Ito · Taira Tsuchiya · Junya Honda
- abstract@[open-review](#): This study considers online learning with general directed feedback graphs. For this problem, we present best-of-both-worlds algorithms that achieve nearly tight regret bounds for adversarial environments as well as poly-logarithmic regret bounds for stochastic environments. As \citet{alon2015online} have shown, tight regret bounds depend on the structure of the feedback graph: \textit{strongly observable} graphs yield minimax regret of  $\tilde{O}(\Theta(\alpha^{1/2} T^{1/2}))$ , while \textit{weakly observable} graphs induce minimax regret of  $\tilde{O}(\Delta^{1/3} T^{2/3})$ , where  $\alpha$  and  $\Delta$ , respectively, represent the independence number of the graph and the domination number of a certain portion of the graph. Our proposed algorithm for strongly observable graphs has a regret bound of  $\tilde{O}(\alpha^{1/2} T^{1/2})$  for adversarial environments, as well as of  $O(\frac{\alpha (\ln T)^3}{\Delta_{\min}})$  for stochastic environments, where  $\Delta_{\min}$  expresses the minimum suboptimality gap. This result resolves an open question raised by \citet{erez2021towards}. We also provide an algorithm for weakly observable graphs that achieves a regret bound of  $\tilde{O}(\Delta^{1/3} T^{2/3})$  for adversarial environments and poly-logarithmic regret for stochastic environments. The proposed algorithms are based on the follow-the-perturbed-leader approach combined with newly designed update rules for learning rates.

## [AutoMS: Automatic Model Selection for Novelty Detection with Error Rate Control](#)

- Yifan Zhang · Haiyan Jiang · Haojie Ren · Changliang Zou · Dejing Dou
- abstract@[open-review](#): Given an unsupervised novelty detection task on a new dataset, how can we automatically select a "best" detection model while simultaneously controlling the error rate of the best model? For novelty detection analysis, numerous detectors have been proposed to detect outliers on a new unseen dataset based on a score function trained on available clean data. However, due to the absence of labeled data for model evaluation and comparison, there is a lack of systematic approaches that are able to select a "best" model/detector (i.e., the algorithm as well as its hyperparameters) and achieve certain error rate control simultaneously. In this paper, we introduce a unified data-driven procedure to address this issue. The key idea is to maximize the number of detected outliers while controlling the false discovery rate (FDR) with the help of Jackknife prediction. We establish non-asymptotic bounds for the false discovery proportions and show that the proposed procedure yields valid FDR control under some mild conditions. Numerical experiments on both synthetic and real data validate the theoretical results and demonstrate the effectiveness of our proposed AutoMS method.

## [Computationally Efficient Horizon-Free Reinforcement Learning for Linear Mixture MDPs](#)

- Dongruo Zhou · Quanquan Gu
- abstract@[open-review](#): Recent studies have shown that episodic reinforcement learning (RL) is not more difficult than bandits, even with a long planning horizon and unknown state transitions. However, these results are limited to either tabular Markov decision processes (MDPs) or computationally inefficient algorithms for linear mixture MDPs. In this paper, we propose the first computationally efficient horizon-free algorithm for linear mixture MDPs, which achieves the optimal  $\tilde{O}(d\sqrt{K} + d^2)$  regret up to logarithmic factors. Our algorithm adapts a weighted least square estimator for the unknown transitional dynamic, where the weight is both \emph{variance-aware} and \emph{uncertainty-aware}. When applying our weighted least square estimator to heterogeneous linear bandits, we can obtain an  $\tilde{O}(d\sqrt{\sum_{k=1}^K \sigma_k^2} + d)$  regret in the first  $K$  rounds, where  $d$  is the dimension of the context and  $\sigma_k^2$  is the variance of the reward in the  $k$ -th round. This also improves upon the best known algorithms in this setting when  $\sigma_k^2$ 's are known.

## [Active Learning Helps Pretrained Models Learn the Intended Task](#)

- Alex Tamkin · Dat Nguyen · Salil Deshpande · Jesse Mu · Noah Goodman
- abstract@[open-review](#): Models can fail in unpredictable ways during deployment due to task ambiguity, when multiple behaviors are consistent with the provided training data. An example is an object classifier trained on red squares and blue circles: when encountering blue squares, the intended behavior is undefined. We investigate whether pretrained models are better active learners, capable of disambiguating between the possible tasks a user may be trying to specify. Intriguingly, we find that better active learning is an emergent property of the pretraining process: pretrained models require up to 5 times fewer labels when using uncertainty-based active learning, while non-pretrained models see no or even negative benefit. We find these gains come from an ability to select examples with attributes that disambiguate the intended behavior, such as rare product categories or atypical backgrounds. These attributes are far more linearly separable in pretrained model's representation spaces vs non-pretrained models, suggesting a possible mechanism for this behavior.

## [Mixture-of-Experts with Expert Choice Routing](#)

- Yanqi Zhou · Tao Lei · Hanxiao Liu · Nan Du · Yanping Huang · Vincent Zhao · Andrew Dai · zhifeng Chen · Quoc V Le · James Laudon
- abstract@[open-review](#): Sparsely-activated Mixture-of-experts (MoE) models allow the number of parameters to greatly increase while keeping the amount of computation for a given token or a given sample unchanged. However, a poor expert routing strategy (e.g. one resulting in load imbalance) can cause certain experts to be under-trained, leading to an expert being under or over-specialized. Prior work allocates a fixed number of experts to each token using a top-k function regardless of the relative importance of different tokens. To address this, we propose a heterogeneous mixture-of-experts employing an expert choice method. Instead of letting tokens select the top-k experts, we have experts selecting the top-k tokens. As a result, each token can be routed to a variable number of experts and each expert can have a fixed bucket size. We systematically study pre-training speedups using the same computational resources of the Switch Transformer top-1 and GShard top-2 gating of prior work and find that our method improves training convergence

time by more than  $2\tilde{A}$ . For the same computational cost, our method demonstrates higher performance in fine-tuning 11 selected tasks in the GLUE and SuperGLUE benchmarks. For a smaller activation cost, our method outperforms the T5 dense model in 7 out of the 11 tasks.

## [Left Heavy Tails and the Effectiveness of the Policy and Value Networks in DNN-based best-first search for Sokoban Planning](#)

- Dieqiao Feng Â· Carla Gomes Â· Bart Selman
- abstract@[open-review](#): Despite the success of practical solvers in various NP-complete domains such as SAT and CSP as well as using deep reinforcement learning to tackle two-player games such as Go, certain classes of PSPACE-hard planning problems have remained out of reach. Even carefully designed domain-specialized solvers can fail quickly due to the exponential search space on hard instances. Recent works that combine traditional search methods, such as best-first search and Monte Carlo tree search, with Deep Neural Networks' (DNN) heuristics have shown promising progress and can solve a significant number of hard planning instances beyond specialized solvers. To better understand why these approaches work, we studied the interplay of the policy and value networks of DNN-based best-first search on Sokoban and show the surprising effectiveness of the policy network, further enhanced by the value network, as a guiding heuristic for the search. To further understand the phenomena, we studied the cost distribution of the search algorithms and found that Sokoban instances can have heavy-tailed runtime distributions, with tails both on the left and right-hand sides. In particular, for the first time, we show the existence of \textit{left heavy tails} and propose an abstract tree model that can empirically explain the appearance of these tails. The experiments show the critical role of the policy network as a powerful heuristic guiding the search, which can lead to left heavy tails with polynomial scaling by avoiding exploring exponentially sized subtrees. Our results also demonstrate the importance of random restarts, as are widely used in traditional combinatorial solvers, for DNN-based search methods to avoid left and right heavy tails.

## [Beyond the Return: Off-policy Function Estimation under User-specified Error-measuring Distributions](#)

- Audrey Huang Â· Nan Jiang
- abstract@[open-review](#): Off-policy evaluation often refers to two related tasks: estimating the expected return of a policy and estimating its value function (or other functions of interest, such as density ratios). While recent works on marginalized importance sampling (MIS) show that the former can enjoy provable guarantees under realizable function approximation, the latter is only known to be feasible under much stronger assumptions such as prohibitively expressive discriminators. In this work, we provide guarantees for off-policy function estimation under only realizability, by imposing proper regularization on the MIS objectives. Compared to commonly used regularization in MIS, our regularizer is much more flexible and can account for an arbitrary user-specified distribution, under which the learned function will be close to the groundtruth. We provide exact characterization of the optimal dual solution that needs to be realized by the discriminator class, which determines the data-coverage assumption in the case of value-function learning. As another surprising observation, the regularizer can be altered to relax the data-coverage requirement, and completely eliminate it in the ideal case with strong side information.

## [MinVIS: A Minimal Video Instance Segmentation Framework without Video-based Training](#)

- De-An Huang Â· Zhiding Yu Â· Anima Anandkumar
- abstract@[open-review](#): We propose MinVIS, a minimal video instance segmentation (VIS) framework that achieves state-of-the-art VIS performance with neither video-based architectures nor training procedures. By only training a query-based image instance segmentation model, MinVIS outperforms the previous best result on the challenging Occluded VIS dataset by over 10% AP. Since MinVIS treats frames in training videos as independent images, we can drastically sub-sample the annotated frames in training videos without any modifications. With only 1% of labeled frames, MinVIS outperforms or is comparable to fully-supervised state-of-the-art approaches on YouTube-VIS 2019/2021. Our key observation is that queries trained to be discriminative between intra-frame object instances are temporally consistent and can be used to track instances without any manually designed heuristics. MinVIS thus has the following inference pipeline: we first apply the trained query-based image instance segmentation to video frames independently. The segmented instances are then tracked by bipartite matching of the corresponding queries. This inference is done in an online fashion and does not need to process the whole video at once. MinVIS thus has the practical advantages of reducing both the labeling costs and the memory requirements, while not sacrificing the VIS performance.

## [Signal Processing for Implicit Neural Representations](#)

- Dejia Xu Â· Peihao Wang Â· Yifan Jiang Â· Zhiwen Fan Â· Zhangyang Wang
- abstract@[open-review](#): Implicit Neural Representations (INR) encoding continuous multi-media data via multi-layer perceptrons has shown undeniable promise in various computer vision tasks. Despite many successful applications, editing and processing an INR remains intractable as signals are represented by agnostic parameters of a neural network. Existing works manipulate such continuous representations via processing on their discretized instance, which breaks down the compactness and continuous nature of INR. In this work, we present a pilot study on the question: how to directly modify an INR without explicit decoding? We answer this question by proposing an implicit neural signal processing network, dubbed INSP-Net, via differential operators on INR. Our key insight is that spatial gradients of neural networks can be computed analytically and invariant to translation, while mathematically we show that any continuous convolution filter can be uniformly approximated by a linear combination of high-order differential operators. With these two knobs, we instantiate the INR signal operator as a composition of computational graphs corresponding to the high-order derivatives, where the weighting parameters can be either handcrafted or data-driven learned. Based on our proposed INSP-Net, we further build the first Convolutional Neural Network (CNN) that implicitly runs on INRs, named INSP-ConvNet. Our experiments validate the expressiveness of INSP-Net and INSP-ConvNet in fitting low-level image processing kernels (e.g. edge detection, blurring, deblurring, denoising, inpainting) as well as for high-level tasks on implicit fields such as image classification. We will release all codes.

## [Scalable Algorithm Synthesis with Recurrent Networks: Extrapolation without Overthinking](#)

- Arpit Bansal Â· Avi Schwarzschild Â· Eitan Borgnia Â· Zeyad Emam Â· Furong Huang Â· Micah Goldblum Â· Tom Goldstein
- abstract@[open-review](#): Machine learning systems perform well on pattern matching tasks, but their ability to perform algorithmic or logical reasoning is not well understood. One important reasoning capability is algorithmic extrapolation, in which models trained only on small/simple reasoning problems can synthesize complex strategies for large/complex problems at test time. Algorithmic extrapolation can be achieved through recurrent systems, which can be iterated many times to solve difficult reasoning problems. We observe that this approach fails to scale to highly complex problems because behavior degenerates when many iterations are applied -- an issue we refer to as "overthinking." We propose a recall architecture that keeps an explicit copy of the problem instance in memory so that it cannot be forgotten. We also employ a progressive training routine that prevents the model from learning behaviors that are specific to iteration number and instead pushes it to learn behaviors that can be repeated indefinitely. These innovations prevent the overthinking problem, and enable recurrent systems to solve extremely hard extrapolation tasks.

## [GraB: Finding Provably Better Data Permutations than Random Reshuffling](#)

- Yucheng Lu Â· Wentao Guo Â· Christopher De Sa
- abstract@[open-review](#): Random reshuffling, which randomly permutes the dataset each epoch, is widely adopted in model training because it yields faster convergence than with-replacement sampling. Recent studies indicate greedily chosen data orderings can further speed up convergence empirically, at the cost of using more computation and memory. However, greedy ordering lacks theoretical justification and has limited utility due to its non-trivial memory and computation overhead. In this paper, we first formulate an example-ordering framework named \textit{herding} and answer affirmatively that SGD with herding converges at the rate  $\$O(T^{-2/3})\$$  on smooth, non-convex objectives, faster than the  $\$O(n^{1/3}T^{-2/3})\$$  obtained by random

reshuffling, where  $n$  denotes the number of data points and  $T$  denotes the total number of iterations. To reduce the memory overhead, we leverage discrepancy minimization theory to propose an online Gradient Balancing algorithm (GraB) that enjoys the same rate as herding, while reducing the memory usage from  $O(nd)$  to just  $O(d)$  and computation from  $O(n^2)$  to  $O(n)$ , where  $d$  denotes the model dimension. We show empirically on applications including MNIST, CIFAR10, WikiText and GLUE that GraB can outperform random reshuffling in terms of both training and validation performance, and even outperform state-of-the-art greedy ordering while reducing memory usage over 100 times.

## [Linear Label Ranking with Bounded Noise](#)

- Dimitris Fotakis · Alkis Kalavasis · Vasilis Kontonis · Christos Tzamos
- abstract@[open-review](#): Label Ranking (LR) is the supervised task of learning a sorting function that maps feature vectors  $x \in \mathbb{R}^d$  to rankings  $\sigma(x) \in \mathbb{S}_k$  over a finite set of  $k$  labels. We focus on the fundamental case of learning linear sorting functions (LSFs) under Gaussian marginals:  $x$  is sampled from the  $d$ -dimensional standard normal and the ground truth ranking  $\sigma^*(x)$  is the ordering induced by sorting the coordinates of the vector  $W^*x$ , where  $W^*$  is unknown. We consider learning LSFs in the presence of bounded noise: assuming that a noiseless example is of the form  $(x, \sigma^*(x))$ , we observe  $(x, \pi)$ , where for any pair of elements  $i \neq j$ , the probability that the order of  $i, j$  is different in  $\pi$  than in  $\sigma^*(x)$  is at most  $\eta < 1/2$ . We design efficient non-proper and proper learning algorithms that learn hypotheses within normalized Kendall's Tau distance  $\epsilon$  from the ground truth with  $N = \tilde{O}(d \log(k) / \epsilon)$  labeled examples and runtime  $\mathrm{poly}(N, k)$ . For the more challenging top-\$r\$ disagreement loss, we give an efficient proper learning algorithm that achieves  $\epsilon$  top-\$r\$ disagreement with the ground truth with  $N = \tilde{O}(d k r / \epsilon)$  samples and  $\mathrm{poly}(N)$  runtime.

## [Re-Analyze Gauss: Bounds for Private Matrix Approximation via Dyson Brownian Motion](#)

- Oren Mangoubi · Nisheeth Vishnoi
- abstract@[open-review](#): Given a symmetric matrix  $M$  and a vector  $\lambda$ , we present new bounds on the Frobenius-distance utility of the Gaussian mechanism for approximating  $M$  by a matrix whose spectrum is  $\lambda$ , under  $(\epsilon, \delta)$ -differential privacy. Our bounds depend on both  $\lambda$  and the gaps in the eigenvalues of  $M$ , and hold whenever the top  $k+1$  eigenvalues of  $M$  have sufficiently large gaps. When applied to the problems of private rank-\$k\$ covariance matrix approximation and subspace recovery, our bounds yield improvements over previous bounds. Our bounds are obtained by viewing the addition of Gaussian noise as a continuous-time matrix Brownian motion. This viewpoint allows us to track the evolution of eigenvalues and eigenvectors of the matrix, which are governed by stochastic differential equations discovered by Dyson. These equations allow us to bound the utility as the square-root of a sum-of-squares of perturbations to the eigenvectors, as opposed to a sum of perturbation bounds obtained via Davis-Kahan-type theorems.

## [FeLMi : Few shot Learning with hard Mixup](#)

- Aniket Roy · Anshul Shah · Ketul Shah · Prithviraj Dhar · Anoop Cherian · Rama Chellappa
- abstract@[open-review](#): Learning from few examples is a challenging computer vision task. Traditionally, meta-learning based methods have been used to solve this problem. However recent approaches show improvements by learning a feature extractor on the abundant base examples and transferring these to the fewer novel examples. However, the finetuning stage is often prone to overfitting due to the small size of the novel dataset. To this end, we propose Few shot Learning with hard Mixup (FeLMi) using manifold mixup to synthetically generate samples which helps in mitigating the data scarcity issue. Different from naive mixup, our approach selects the hard mixup samples using an uncertainty based criteria. To the best of our knowledge, we are the first to use hard-mixup for few-shot learning problem. Our approach allows to better exploit the pseudo-labeled base examples through base-novel mixup and entropy based filtering. We evaluate our approach on common few-shot benchmarks, e.g., FC-100, CIFAR-FS and miniImageNet and obtain improvement in both 1-shot and 5-shot settings.

## [Inherently Explainable Reinforcement Learning in Natural Language](#)

- Xiangyu Peng · Mark Riedl · Prithviraj Ammanabrolu
- abstract@[open-review](#): We focus on the task of creating a reinforcement learning agent that is inherently explainable---with the ability to produce immediate local explanations by thinking out loud while performing a task and analyzing entire trajectories post-hoc to produce temporally extended explanations. This Hierarchically Explainable Reinforcement Learning agent (HEX-RL), operates in Interactive Fictions, text-based game environments in which an agent perceives and acts upon the world using textual natural language. These games are usually structured as puzzles or quests with long-term dependencies in which an agent must complete a sequence of actions to succeed---providing ideal environments in which to test an agent's ability to explain its actions. Our agent is designed to treat explainability as a first-class citizen, using an extracted symbolic knowledge graph-based state representation coupled with a Hierarchical Graph Attention mechanism that points to the facts in the internal graph representation that most influenced the choice of actions. Experiments show that this agent provides significantly improved explanations over strong baselines, as rated by human participants generally unfamiliar with the environment, while also matching state-of-the-art task performance.

## [Unknown-Aware Domain Adversarial Learning for Open-Set Domain Adaptation](#)

- JoonHo Jang · Byeonghu Na · Dong Hyeok Shin · Mingi Ji · Kyungwoo Song · Il-chul Moon
- abstract@[open-review](#): Open-Set Domain Adaptation (OSDA) assumes that a target domain contains unknown classes, which are not discovered in a source domain. Existing domain adversarial learning methods are not suitable for OSDA because distribution matching with  $\text{unknown}$  classes leads to negative transfer. Previous OSDA methods have focused on matching the source and the target distribution by only utilizing  $\text{known}$  classes. However, this  $\text{known}$ -only matching may fail to learn the target- $\text{unknown}$  feature space. Therefore, we propose Unknown-Aware Domain Adversarial Learning (UADAL), which aligns the source and the target- $\text{known}$  distribution while simultaneously segregating the target- $\text{unknown}$  distribution in the feature alignment procedure. We provide theoretical analyses on the optimized state of the proposed  $\text{unknown-aware}$  feature alignment, so we can guarantee both  $\text{alignment}$  and  $\text{segregation}$  theoretically. Empirically, we evaluate UADAL on the benchmark datasets, which shows that UADAL outperforms other methods with better feature alignments by reporting state-of-the-art performances.

## [Improving Variational Autoencoders with Density Gap-based Regularization](#)

- Jianfei Zhang · Jun Bai · Chenghua Lin · Yanmeng Wang · Wenge Rong
- abstract@[open-review](#): Variational autoencoders (VAEs) are one of the powerful unsupervised learning frameworks in NLP for latent representation learning and latent-directed generation. The classic optimization goal of VAEs is to maximize the Evidence Lower Bound (ELBO), which consists of a conditional likelihood for generation and a negative Kullback-Leibler (KL) divergence for regularization. In practice, optimizing ELBO often leads the posterior distribution of all samples converge to the same degenerated local optimum, namely posterior collapse or KL vanishing. There are effective ways proposed to prevent posterior collapse in VAEs, but we observe that they in essence make trade-offs between posterior collapse and hole problem, i.e., mismatch between the aggregated posterior distribution and the prior distribution. To this end, we introduce new training objectives to tackle both two problems through a novel regularization based on the probabilistic density gap between the aggregated posterior distribution and the prior distribution. Through experiments on language modeling, latent space visualization and interpolation, we show that our proposed method can solve both problems

effectively and thus outperforms the existing methods in latent-directed generation. To the best of our knowledge, we are the first to jointly solve the hole problem and the posterior collapse.

## [What's the Harm? Sharp Bounds on the Fraction Negatively Affected by Treatment](#)

- Nathan Kallus
- abstract@[open-review](#): The fundamental problem of causal inference -- that we never observe counterfactuals -- prevents us from identifying how many might be negatively affected by a proposed intervention. If, in an A/B test, half of users click (or buy, or watch, or renew, etc.), whether exposed to the standard experience A or a new one B, hypothetically it could be because the change affects no one, because the change positively affects half the user population to go from no-click to click while negatively affecting the other half, or something in between. While unknowable, this impact is clearly of material importance to the decision to implement a change or not, whether due to fairness, long-term, systemic, or operational considerations. We therefore derive the tightest-possible (i.e., sharp) bounds on the fraction negatively affected (and other related estimands) given data with only factual observations, whether experimental or observational. Naturally, the more we can stratify individuals by observable covariates, the tighter the sharp bounds. Since these bounds involve unknown functions that must be learned from data, we develop a robust inference algorithm that is efficient almost regardless of how and how fast these functions are learned, remains consistent when some are mislearned, and still gives valid conservative bounds when most are mislearned. Our methodology altogether therefore strongly supports credible conclusions: it avoids spuriously point-identifying this unknowable impact, focusing on the best bounds instead, and it permits exceedingly robust inference on these. We demonstrate our method in simulation studies and in a case study of career counseling for the unemployed.

## [Thor: Wielding Hammers to Integrate Language Models and Automated Theorem Provers](#)

- Albert Qiaochu Jiang · Wenda Li · Szymon Tworkowski · Konrad Czechowski · Tomasz Odrzygowski · Piotr Miśoń · Yuhuai Wu · Mateja Jamnik
- abstract@[open-review](#): In theorem proving, the task of selecting useful premises from a large library to unlock the proof of a given conjecture is crucially important. This presents a challenge for all theorem provers, especially the ones based on language models, due to their relative inability to reason over huge volumes of premises in text form. This paper introduces Thor, a framework integrating language models and automated theorem provers to overcome this difficulty. In Thor, a class of methods called hammers that leverage the power of automated theorem provers are used for premise selection, while all other tasks are designated to language models. Thor increases a language model's success rate on the PISA dataset from \$39\%\$ to \$57\%\$, while solving \$8.2\%\$ of problems neither language models nor automated theorem provers are able to solve on their own. Furthermore, with a significantly smaller computational budget, Thor can achieve a success rate on the MiniF2F dataset that is on par with the best existing methods. Thor can be instantiated for the majority of popular interactive theorem provers via a straightforward protocol we provide.

## [Diagnosing failures of fairness transfer across distribution shift in real-world medical settings](#)

- Jessica Schrouff · Natalie Harris · Sanmi Koyejo · Ibrahim Alabdulmohsin · Eva Schnider · Krista Opsahl-Ong · Alexander Brown · Subhrajit Roy · Diana Mincu · Christina Chen · Awa Dieng · Yuan Liu · Vivek Natarajan · Alan Karthikesalingam · Katherine Heller · Silvia Chiappa · Alexander D'Amour
- abstract@[open-review](#): Diagnosing and mitigating changes in model fairness under distribution shift is an important component of the safe deployment of machine learning in healthcare settings. Importantly, the success of any mitigation strategy strongly depends on the \textit{structure} of the shift. Despite this, there has been little discussion of how to empirically assess the structure of a distribution shift that one is encountering in practice. In this work, we adopt a causal framing to motivate conditional independence tests as a key tool for characterizing distribution shifts. Using our approach in two medical applications, we show that this knowledge can help diagnose failures of fairness transfer, including cases where real-world shifts are more complex than is often assumed in the literature. Based on these results, we discuss potential remedies at each step of the machine learning pipeline.

## [ZooD: Exploiting Model Zoo for Out-of-Distribution Generalization](#)

- Qishi Dong · Awais Muhammad · Fengwei Zhou · Chuanlong Xie · Tianyang Hu · Yongxin Yang · Sung-Ho Bae · Zhenguo Li
- abstract@[open-review](#): Recent advances on large-scale pre-training have shown great potentials of leveraging a large set of Pre-Trained Models (PTMs) for improving Out-of-Distribution (OoD) generalization, for which the goal is to perform well on possible unseen domains after fine-tuning on multiple training domains. However, maximally exploiting a zoo of PTMs is challenging since fine-tuning all possible combinations of PTMs is computationally prohibitive while accurate selection of PTMs requires tackling the possible data distribution shift for OoD tasks. In this work, we propose ZooD, a paradigm for PTMs ranking and ensemble with feature selection. Our proposed metric ranks PTMs by quantifying inter-class discriminability and inter-domain stability of the task data features extracted by the PTMs in a leave-one-domain-out cross-validation manner. The top-K ranked models are then aggregated for the target OoD task. To avoid accumulating noise induced by model ensemble, we propose an efficient variational EM algorithm to select informative features. We evaluate our paradigm on a diverse model zoo consisting of 35 models for various OoD tasks and demonstrate: (i) model ranking is better correlated with fine-tuning ranking than previous methods and up to 9859x faster than brute-force fine-tuning; (ii) OoD generalization outperforms the state-of-the-art methods and accuracy on most challenging task DomainNet is improved from 46.5% to 50.6%.

## [SU-SSL: Maximize Performance in Unseen Classes and Maintain Safeness in Seen Classes](#)

- Yi-Ge Zhang · Lan-Zhe Guo · Zhi-Fan Wu · Jie-Jing Shao · Yu-Feng Li
- abstract@[open-review](#): Semi-supervised learning (SSL) has received tremendous attention due to its ability to leverage unlabeled data. Existing SSL methods typically assume all unlabeled data are from seen classes, i.e., classes are observed in the labeled dataset. However, in real-world applications, unseen classes are commonly occurred, which severely degrade SSL performance on seen classes. Open-set SSL methods are designed to maintain safeness on seen classes, but they fail to classify unseen classes. Novel class discovery (NCD) methods aim to discover unseen classes automatically but it is unsafe for seen classes. In this paper, we develop a new SSL approach, called Safe Unseen classification Semi-Supervised Learning (SU-SSL), which can not only classify unseen classes automatically but also maintain safeness on seen classes. Our approach consists of two modules: Unseen Class Classification and Adaptive Threshold. Specifically, we first improve the SSL methods to discover unseen classes by proposing a new unseen class classification objective that can exploit pairwise similarity and eliminate potential noisy pairs, and then bridge the performance gap between seen and unseen classes by proposing an adaptive threshold based SSL objective. Extensive empirical evaluations show our approach achieves 37.7% improvement in unseen-class classification compared with SSL methods, and 26.3% improvement in seen class compared with NCD methods.

## [Variable Experience Rollout: Training Robust Skill Policies for Rearrangement](#)

- Erik Wijmans · Irfan Essa · Dhruv Batra
- abstract@[open-review](#): We present Variable Experience Rollout (VER), a technique for efficiently scaling batched on-policy reinforcement learning in heterogeneous environments (where different environments take vastly different times for generating rollouts) to many GPUs residing on, potentially, many machines. VER combines the strengths of and blurs the line between synchronous and asynchronous on-policy RL methods (SyncOnRL and AsyncOnRL, respectively). Specifically, it learns from on-policy experience (like SyncOnRL) and has no synchronization points (like AsyncOnRL), enabling high throughput. We find that VER leads to significant and consistent speed-ups across a broad range of embodied navigation and mobile manipulation tasks in photorealistic 3D simulation environments. Specifically, for PointGoal navigation and ObjectGoal navigation in Habitat 1.0, VER is 60-100% faster (1.6-2x speedup) than DD-PPO, the current state of art for distributed SyncOnRL, with similar sample efficiency. For mobile manipulation tasks (open

fridge/cabinet, pick/place objects) in Habitat 2.0 VER is 150% faster (2.5x speedup) on 1 GPU and 170% faster (2.7x speedup) on 8 GPUs than DD-PPO. Compared to SampleFactory (the current state-of-the-art AsyncOnRL), VER matches its speed on 1 GPU, and is 70% faster (1.7x speedup) on 8 GPUs with better sample efficiency. We leverage these speed-ups to train chained skills for GeometricGoal rearrangement tasks in the Home Assistant Benchmark (HAB). We find a surprising emergence of navigation in skills that do not ostensibly require any navigation. Specifically, the Pick skill involves a robot picking an object from a table. During training the robot was always spawned close to the table and never needed to navigate. However, we find that if base movement is part of the action space, the robot learns to navigate then pick an object in new environments with 50% success, demonstrating surprisingly high out-of-distribution generalization.

## [Polynomial-Time Optimal Equilibria with a Mediator in Extensive-Form Games](#)

- Brian Zhang · Tuomas Sandholm
- abstract@[open-review](#): For common notions of correlated equilibrium in extensive-form games, computing an optimal (e.g., welfare-maximizing) equilibrium is NP-hard. Other equilibrium notions---communication and certification equilibria---augment the game with a mediator that has the power to both send and receive messages to and from the players---and, in particular, to remember the messages. In this paper, we investigate both notions in extensive-form games from a computational lens. We show that optimal equilibria in both notions can be computed in polynomial time, the latter under a natural additional assumption known in the literature. Our proof works by constructing a {\em mediator-augmented game} of polynomial size that explicitly represents the mediator's decisions and actions. Our framework allows us to define an entire family of equilibria by varying the mediator's information partition, the players' ability to lie, and the players' ability to deviate. From this perspective, we show that other notions of equilibrium, such as extensive-form correlated equilibrium, correspond to the mediator having imperfect recall. This shows that, at least among all these equilibrium notions, the hardness of computation is driven by the mediator's imperfect recall. As special cases of our general construction, we recover the polynomial-time algorithm of Conitzer & Sandholm [2004] for automated mechanism design in Bayes-Nash equilibria, and the correlation DAG algorithm of Zhang et al [2022] for optimal correlation. Our algorithm is especially scalable when the equilibrium notion is what we define as the full-certification equilibrium, where players cannot lie about their information but they can be silent. We back up our theoretical claims with experiments on a suite of standard benchmark games.

## [Verification and search algorithms for causal DAGs](#)

- Davin Choo · Kirankumar Shiragur · Arnab Bhattacharyya
- abstract@[open-review](#): We study two problems related to recovering causal graphs from interventional data: (i) \textsf{verification}, where the task is to check if a purported causal graph is correct, and (ii) \textsf{search}, where the task is to recover the correct causal graph. For both, we wish to minimize the number of interventions performed. For the first problem, we give a characterization of a minimal sized set of atomic interventions that is necessary and sufficient to check the correctness of a claimed causal graph. Our characterization uses the notion of \textsf{covered edges}, which enables us to obtain simple proofs and also easily reason about earlier results. We also generalize our results to the settings of bounded size interventions and node-dependent interventional costs. For all the above settings, we provide the first known provable algorithms for efficiently computing (near)-optimal verifying sets on general graphs. For the second problem, we give a simple adaptive algorithm based on graph separators that produces an atomic intervention set which fully orients any essential graph while using  $\mathcal{O}(\log n)$  times the optimal number of interventions needed to \textsf{verify} (verifying size) the underlying DAG on  $n$  vertices. This approximation is tight as \textsf{any} search algorithm on an essential line graph has worst case approximation ratio of  $\Omega(\log n)$  with respect to the verifying size. With bounded size interventions, each of size  $\leq k$ , our algorithm gives an  $\mathcal{O}(\log n \cdot \log \log k)$  factor approximation. Our result is the first known algorithm that gives a non-trivial approximation guarantee to the verifying size on general unweighted graphs and with bounded size interventions.

## [A Quadrature Rule combining Control Variates and Adaptive Importance Sampling](#)

- Romain Leluc · François Portier · Johan Segers · Aigerim Zhuman
- abstract@[open-review](#): Driven by several successful applications such as in stochastic gradient descent or in Bayesian computation, control variates have become a major tool for Monte Carlo integration. However, standard methods do not allow the distribution of the particles to evolve during the algorithm, as is the case in sequential simulation methods. Within the standard adaptive importance sampling framework, a simple weighted least squares approach is proposed to improve the procedure with control variates. The procedure takes the form of a quadrature rule with adapted quadrature weights to reflect the information brought in by the control variates. The quadrature points and weights do not depend on the integrand, a computational advantage in case of multiple integrands. Moreover, the target density needs to be known only up to a multiplicative constant. Our main result is a non-asymptotic bound on the probabilistic error of the procedure. The bound proves that for improving the estimate's accuracy, the benefits from adaptive importance sampling and control variates can be combined. The good behavior of the method is illustrated empirically on synthetic examples and real-world data for Bayesian linear regression.

## [Tempo: Accelerating Transformer-Based Model Training through Memory Footprint Reduction](#)

- Muralidhar Andoorveedu · Zhanda Zhu · Bojian Zheng · Gennady Pekhimenko
- abstract@[open-review](#): Training deep learning models can be computationally expensive. Prior works have shown that increasing the batch size can potentially lead to better overall throughput. However, the batch size is frequently limited by the accelerator memory capacity due to the activations/feature maps stored for the training backward pass, as larger batch sizes require larger feature maps to be stored. Transformer-based models, which have recently seen a surge in popularity due to their good performance and applicability to a variety of tasks, have a similar problem. To remedy this issue, we propose Tempo, a new approach to efficiently use accelerator (e.g., GPU) memory resources for training Transformer-based models. Our approach provides drop-in replacements for the GELU, LayerNorm, and Attention layers, reducing the memory usage and ultimately leading to more efficient training. We implement Tempo and evaluate the throughput, memory usage, and accuracy/loss on the BERT Large pre-training task. We demonstrate that Tempo enables up to 2x higher batch sizes and 16% higher training throughput over the state-of-the-art baseline. We also evaluate Tempo on GPT2 and RoBERTa models, showing 19% and 26% speedup over the baseline.

## [INRAS: Implicit Neural Representation for Audio Scenes](#)

- Kun Su · Mingfei Chen · Eli Shlizerman
- abstract@[open-review](#): The spatial acoustic information of a scene, i.e., how sounds emitted from a particular location in the scene are perceived in another location, is key for immersive scene modeling. Robust representation of scene's acoustics can be formulated through a continuous field formulation along with impulse responses varied by emitter-listener locations. The impulse responses are then used to render sounds perceived by the listener. While such representation is advantageous, parameterization of impulse responses for generic scenes presents itself as a challenge. Indeed, traditional acoustic field coding methods only implement parameterization at discrete probe points and rely on handcrafted features. In this work, we introduce a novel method for Implicit Neural Representation for Audio Scenes (INRAS) which renders high fidelity time-domain impulse responses at any arbitrary emitter-listener positions using neural network parameterization. Our experimental results show that INRAS outperforms existing approaches for the representation and rendering of sounds for varying emitter-listener locations in all aspects, including the impulse response quality, inference speed, and storage requirements. INRAS achieves such enhancement in performance by introducing a novel audio scene feature decomposition, which leads to efficient reuse of scene-dependent features for any arbitrary emitter-listener positions. Furthermore, such a decomposition allows INRAS to generalize the representation from one scene to another with only a few additional parameters.

## [Training and Inference on Any-Order Autoregressive Models the Right Way](#)

- Andy Shih · Dorsa Sadigh · Stefano Ermon
- abstract@[open-review](#): Conditional inference on arbitrary subsets of variables is a core problem in probabilistic inference with important applications such as masked language modeling and image inpainting. In recent years, the family of Any-Order Autoregressive Models (AO-ARMs) -- which includes popular models such as XLNet -- has shown breakthrough performance in arbitrary conditional tasks across a sweeping range of domains. But, in spite of their success, in this paper we identify significant improvements to be made to previous formulations of AO-ARMs. First, we show that AO-ARMs suffer from redundancy in their probabilistic model, i.e., they define the same distribution in multiple different ways. We alleviate this redundancy by training on a smaller set of univariate conditionals that still maintains support for efficient arbitrary conditional inference. Second, we upweight the training loss for univariate conditionals that are evaluated more frequently during inference. Our method leads to improved performance with no compromises on tractability, giving state-of-the-art likelihoods in arbitrary conditional modeling on text (Text8), image (CIFAR10, ImageNet32), and continuous tabular data domains.

## [Learning Equivariant Segmentation with Instance-Unique Querying](#)

- Wenguan Wang · James Liang · Dongfang Liu
- abstract@[open-review](#): Prevalent state-of-the-art instance segmentation methods fall into a query-based scheme, in which instance masks are derived by querying the image feature using a set of instance-aware embeddings. In this work, we devise a new training framework that boosts query-based models through discriminative query embedding learning. It explores two essential properties, namely dataset-level uniqueness and transformation equivariance, of the relation between queries and instances. First, our algorithm uses the queries to retrieve the corresponding instances from the whole training dataset, instead of only searching within individual scenes. As querying instances across scenes is more challenging, the segmenters are forced to learn more instance-unique queries for effective instance separation. Second, our algorithm encourages both image (instance) representations and queries to be equivariant against geometric transformations, leading to more robust, instance-query matching. We experimentally show, on top of four famous, query-based models (i.e., CondInst, SOLOv2, SOTR, and Mask2Former), our algorithm provides solid performance gains on COCO dataset. Our code will be released.

## [Improving Policy Learning via Language Dynamics Distillation](#)

- Victor Zhong · Jesse Mu · Luke Zettlemoyer · Edward Grefenstette · Tim Rocktäschel
- abstract@[open-review](#): Recent work has shown that augmenting environments with language descriptions improves policy learning. However, for environments with complex language abstractions, learning how to ground language to observations is difficult due to sparse, delayed rewards. We propose Language Dynamics Distillation (LDD), which pretrains a model to predict environment dynamics given demonstrations with language descriptions, and then fine-tunes these language-aware pretrained representations via reinforcement learning (RL). In this way, the model is trained to both maximize expected reward and retain knowledge about how language relates to environment dynamics. On SILG, a benchmark of five tasks with language descriptions that evaluate distinct generalization challenges on unseen environments (NetHack, ALFWorld, RTFM, Messenger, and Touchdown), LDD outperforms tabula-rasa RL, VAE pretraining, and methods that learn from unlabeled demonstrations in inverse RL and reward shaping with pretrained experts. In our analyses, we show that language descriptions in demonstrations improve sample-efficiency and generalization across environments, and that dynamics modeling with expert demonstrations is more effective than with non-experts.

## [In Defense of the Unitary Scalarization for Deep Multi-Task Learning](#)

- Vitaly Kurin · Alessandro De Palma · Ilya Kostrikov · Shimon Whiteson · Pawan K Mudigonda
- abstract@[open-review](#): Recent multi-task learning research argues against unitary scalarization, where training simply minimizes the sum of the task losses. Several ad-hoc multi-task optimization algorithms have instead been proposed, inspired by various hypotheses about what makes multi-task settings difficult. The majority of these optimizers require per-task gradients, and introduce significant memory, runtime, and implementation overhead. We show that unitary scalarization, coupled with standard regularization and stabilization techniques from single-task learning, matches or improves upon the performance of complex multi-task optimizers in popular supervised and reinforcement learning settings. We then present an analysis suggesting that many specialized multi-task optimizers can be partly interpreted as forms of regularization, potentially explaining our surprising results. We believe our results call for a critical reevaluation of recent research in the area.

## [Recovering Private Text in Federated Learning of Language Models](#)

- Samyak Gupta · Yangsibo Huang · Zexuan Zhong · Tianyu Gao · Kai Li · Danqi Chen
- abstract@[open-review](#): Federated learning allows distributed users to collaboratively train a model while keeping each user's data private. Recently, a growing body of work has demonstrated that an eavesdropping attacker can effectively recover image data from gradients transmitted during federated learning. However, little progress has been made in recovering text data. In this paper, we present a novel attack method FILM (Federated Inversion attack for Language Models) for federated learning of language models---for the first time, we show the feasibility of recovering text from large batch sizes of up to 128 sentences. Different from image-recovery methods which are optimized to match gradients, we take a distinct approach that first identifies a set of words from gradients and then directly reconstructs sentences based on beam search and a prior-based reordering strategy. The key insight of our attack is to leverage either prior knowledge in pre-trained language models or memorization during training. Despite its simplicity, we demonstrate that FILM can work well with several large-scale datasets---it can extract single sentences with high fidelity even for large batch sizes and recover multiple sentences from the batch successfully if the attack is applied iteratively. We hope our results can motivate future work in developing stronger attacks as well as new defense methods for training language models in federated learning.

## [Generalised Mutual Information for Discriminative Clustering](#)

- Louis Ohl · Pierre-Alexandre Mattei · Charles Bouveyron · Warith Harchaoui · Arnaud Droit · Mickaël Leclercq · Frederic Precioso
- abstract@[open-review](#): In the last decade, recent successes in deep clustering majorly involved the mutual information (MI) as an unsupervised objective for training neural networks with increasing regularisations. While the quality of the regularisations have been largely discussed for improvements, little attention has been dedicated to the relevance of MI as a clustering objective. In this paper, we first highlight how the maximisation of MI does not lead to satisfying clusters. We identified the Kullback-Leibler divergence as the main reason of this behaviour. Hence, we generalise the mutual information by changing its core distance, introducing the generalised mutual information (GEMINI): a set of metrics for unsupervised neural network training. Unlike MI, some GEMINIs do not require regularisations when training. Some of these metrics are geometry-aware thanks to distances or kernels in the data space. Finally, we highlight that GEMINIs can automatically select a relevant number of clusters, a property that has been little studied in deep clustering context where the number of clusters is a priori unknown.

## [In the Eye of the Beholder: Robust Prediction with Causal User Modeling](#)

- Amir Feder · Guy Horowitz · Yoav Wald · Roi Reichart · Nir Rosenfeld
- abstract@[open-review](#): Accurately predicting the relevance of items to users is crucial to the success of many social platforms. Conventional approaches train models on logged historical data; but recommendation systems, media services, and online marketplaces all exhibit a constant influx of new content--

-making relevancy a moving target, to which standard predictive models are not robust. In this paper, we propose a learning framework for relevance prediction that is robust to changes in the data distribution. Our key observation is that robustness can be obtained by accounting for how users causally perceive the environment}. We model users as boundedly-rational decision makers whose causal beliefs are encoded by a causal graph, and show how minimal information regarding the graph can be used to contend with distributional changes. Experiments in multiple settings demonstrate the effectiveness of our approach.

## [Why So Pessimistic? Estimating Uncertainties for Offline RL through Ensembles, and Why Their Independence Matters](#)

- Seyed Kamyar Seyed Ghasemipour · Shixiang (Shane) Gu · Ofir Nachum
- abstract@[open-review](#): Motivated by the success of ensembles for uncertainty estimation in supervised learning, we take a renewed look at how ensembles of  $Q$ -functions can be leveraged as the primary source of pessimism for offline reinforcement learning (RL). We begin by identifying a critical flaw in a popular algorithmic choice used by many ensemble-based RL algorithms, namely the use of shared pessimistic target values when computing each ensemble member's Bellman error. Through theoretical analyses and construction of examples in toy MDPs, we demonstrate that shared pessimistic targets can paradoxically lead to value estimates that are effectively optimistic. Given this result, we propose MSG, a practical offline RL algorithm that trains an ensemble of  $Q$ -functions with independently computed targets based on completely separate networks, and optimizes a policy with respect to the lower confidence bound of predicted action values. Our experiments on the popular D4RL and RL Unplugged offline RL benchmarks demonstrate that on challenging domains such as antmazes, MSG with deep ensembles surpasses highly well-tuned state-of-the-art methods by a wide margin. Additionally, through ablations on benchmarks domains, we verify the critical significance of using independently trained  $Q$ -functions, and study the role of ensemble size. Finally, as MSG's use of separate networks per ensemble member can become computationally costly with larger neural network architectures, we investigate whether efficient ensemble approximations developed for supervised learning can be similarly effective, and demonstrate that they do not match the performance and robustness of MSG, highlighting the need for new efforts into efficient uncertainty estimation directed at RL.

## [GMMSeg: Gaussian Mixture based Generative Semantic Segmentation Models](#)

- Chen Liang · Wenguan Wang · Jiaxu Miao · Yi Yang
- abstract@[open-review](#): Prevalent semantic segmentation solutions are, in essence, a dense discriminative classifier of  $p(\text{class}|\text{pixel feature})$ . Though straightforward, this de facto paradigm neglects the underlying data distribution  $p(\text{pixel feature}|\text{class})$ , and struggles to identify out-of-distribution data. Going beyond this, we propose GMMSeg, a new family of segmentation models that rely on a dense generative classifier for the joint distribution  $p(\text{pixel feature}, \text{class})$ . For each class, GMMSeg builds Gaussian Mixture Models (GMMs) via Expectation-Maximization (EM), so as to capture class-conditional densities. Meanwhile, the deep dense representation is end-to-end trained in a discriminative manner, i.e., maximizing  $p(\text{class}|\text{pixel feature})$ . This endows GMMSeg with the strengths of both generative and discriminative models. With a variety of segmentation architectures and backbones, GMMSeg outperforms the discriminative counterparts on three closed-set datasets. More impressively, without any modification, GMMSeg even performs well on open-world datasets. We believe this work brings fundamental insights. Our implementation will be released.

## [Reinforcement Learning with Automated Auxiliary Loss Search](#)

- Tairan He · Yuge Zhang · Kan Ren · Che Wang · Minghuan Liu · Weinan Zhang · Dongsheng Li · Yuqing Yang
- abstract@[open-review](#): A good state representation is crucial to solving complicated reinforcement learning (RL) challenges. Many recent works focus on designing auxiliary losses for learning informative representations. Unfortunately, these handcrafted objectives rely heavily on expert knowledge and may be sub-optimal. In this paper, we propose a principled and universal method for learning better representations with auxiliary loss functions, named Automated Auxiliary Loss Search (A2LS), which automatically searches for top-performing auxiliary loss functions for RL. Specifically, based on the collected trajectory data, we define a general auxiliary loss space of size  $7.5 \times 10^{20}$  and explore the space with an efficient evolutionary search strategy. Empirical results show that the discovered auxiliary loss (namely, A2-winner) significantly improves the performance on both high-dimensional (image) and low-dimensional (vector) unseen tasks with much higher efficiency, showing promising generalization ability to different settings and even different benchmark domains. We conduct a statistical analysis to reveal the relations between patterns of auxiliary losses and RL performance.

## [Learning Neural Acoustic Fields](#)

- Andrew Luo · Yilun Du · Michael Tarr · Josh Tenenbaum · Antonio Torralba · Chuang Gan
- abstract@[open-review](#): Our environment is filled with rich and dynamic acoustic information. When we walk into a cathedral, the reverberations as much as appearance inform us of the sanctuary's wide open space. Similarly, as an object moves around us, we expect the sound emitted to also exhibit this movement. While recent advances in learned implicit functions have led to increasingly higher quality representations of the visual world, there have not been commensurate advances in learning spatial auditory representations. To address this gap, we introduce Neural Acoustic Fields (NAFs), an implicit representation that captures how sounds propagate in a physical scene. By modeling acoustic propagation in a scene as a linear time-invariant system, NAFs learn to continuously map all emitter and listener location pairs to a neural impulse response function that can then be applied to arbitrary sounds. We demonstrate that the continuous nature of NAFs enables us to render spatial acoustics for a listener at an arbitrary location, and can predict sound propagation at novel locations. We further show that the representation learned by NAFs can help improve visual learning with sparse views. Finally, we show that a representation informative of scene structure emerges during the learning of NAFs.

## [Multi-Agent Reinforcement Learning is a Sequence Modeling Problem](#)

- Muning Wen · Jakub Kuba · Runji Lin · Weinan Zhang · Ying Wen · Jun Wang · Yaodong Yang
- abstract@[open-review](#): Large sequence models (SM) such as GPT series and BERT have displayed outstanding performance and generalization capabilities in natural language process, vision and recently reinforcement learning. A natural follow-up question is how to abstract multi-agent decision making also as an sequence modeling problem and benefit from the prosperous development of the SMs. In this paper, we introduce a novel architecture named Multi-Agent Transformer (MAT) that effectively casts cooperative multi-agent reinforcement learning (MARL) into SM problems wherein the objective is to map agents' observation sequences to agents' optimal action sequences. Our goal is to build the bridge between MARL and SMs so that the modeling power of modern sequence models can be unleashed for MARL. Central to our MAT is an encoder-decoder architecture which leverages the multi-agent advantage decomposition theorem to transform the joint policy search problem into a sequential decision making process; this renders only linear time complexity for multi-agent problems and, most importantly, endows MAT with monotonic performance improvement guarantee. Unlike prior arts such as Decision Transformer fit only pre-collected offline data, MAT is trained by online trial and error from the environment in an on-policy fashion. To validate MAT, we conduct extensive experiments on StarCraftII, Multi-Agent MuJoCo, Dexterous Hands Manipulation, and Google Research Football benchmarks. Results demonstrate that MAT achieves superior performance and data efficiency compared to strong baselines including MAPPO and HAPPO. Furthermore, we demonstrate that MAT is an excellent few-short learner on unseen tasks regardless of changes in the number of agents. See our project page at <https://sites.google.com/view/multi-agent-transformer>.

## [Robust Continual Test-time Adaptation: Instance-aware BN and Prediction-balanced Memory](#)

- Taesik Gong · Jongheon Jeong · Taewon Kim · Yewon Kim · Jinwoo Shin · Sung-Ju Lee
- abstract@[open-review](#): Test-time adaptation (TTA) is an emerging paradigm that addresses distributional shifts between training and testing phases without additional data acquisition or labeling cost; only unlabeled test data streams are used for continual model adaptation. Previous TTA schemes

assume that the test samples are independent and identically distributed (i.i.d.), even though they are often temporally correlated (non-i.i.d.) in application scenarios, e.g., autonomous driving. We discover that most existing TTA methods fail dramatically under such scenarios. Motivated by this, we present a new test-time adaptation scheme that is robust against non-i.i.d. test data streams. Our novelty is mainly two-fold: (a) Instance-Aware Batch Normalization (IABN) that corrects normalization for out-of-distribution samples, and (b) Prediction-balanced Reservoir Sampling (PBRs) that simulates i.i.d. data stream from non-i.i.d. stream in a class-balanced manner. Our evaluation with various datasets, including real-world non-i.i.d. streams, demonstrates that the proposed robust TTA not only outperforms state-of-the-art TTA algorithms in the non-i.i.d. setting, but also achieves comparable performance to those algorithms under the i.i.d. assumption.

## [Simultaneous Missing Value Imputation and Structure Learning with Groups](#)

- Pablo Morales-Alvarez · Wenbo Gong · Angus Lamb · Simon Woodhead · Simon Peyton Jones · Nick Pawlowski · Miltiadis Allamanis · Cheng Zhang
- abstract@[open-review](#): Learning structures between groups of variables from data with missing values is an important task in the real world, yet difficult to solve. One typical scenario is discovering the structure among topics in the education domain to identify learning pathways. Here, the observations are student performances for questions under each topic which contain missing values. However, most existing methods focus on learning structures between a few individual variables from the complete data. In this work, we propose VISL, a novel scalable structure learning approach that can simultaneously infer structures between groups of variables under missing data and perform missing value imputations with deep learning. Particularly, we propose a generative model with a structured latent space and a graph neural network-based architecture, scaling to a large number of variables. Empirically, we conduct extensive experiments on synthetic, semi-synthetic, and real-world education data sets. We show improved performances on both imputation and structure learning accuracy compared to popular and recent approaches.

## [Mining Unseen Classes via Regional Objectness: A Simple Baseline for Incremental Segmentation](#)

- Zekang Zhang · Zekang Zhang · Yunchao Wei · Zhiyuan Fang · Jianbo Jiao
- abstract@[open-review](#): Incremental or continual learning has been extensively studied for image classification tasks to alleviate catastrophic forgetting, a phenomenon in which earlier learned knowledge is forgotten when learning new concepts. For class incremental semantic segmentation, such a phenomenon often becomes much worse due to the semantic shift of the background class, i.e., some concepts learned at previous stages are assigned to the background class at the current training stage, therefore, significantly reducing the performance of these old concepts. To address this issue, we propose a simple yet effective method in this paper, named Mining unseen Classes via Regional Objectness (MicroSeg). Our MicroSeg is based on the assumption that \{background regions with strong objectness possibly belong to those concepts in the historical or future stages\}. Therefore, to avoid forgetting old knowledge at the current training stage, our MicroSeg first splits the given image into hundreds of segment proposals with a proposal generator. Those segment proposals with strong objectness from the background are then clustered and assigned new defined labels during the optimization. In this way, the distribution characterizes of old concepts in the feature space could be better perceived, relieving the catastrophic forgetting caused by the semantic shift of the background class accordingly. We conduct extensive experiments on Pascal VOC and ADE20K, and competitive results well demonstrate the effectiveness of our MicroSeg. Code is available in supplementary materials.

## [Exponential Separations in Symmetric Neural Networks](#)

- Aaron Zweig · Joan Bruna
- abstract@[open-review](#): In this work we demonstrate a novel separation between symmetric neural network architectures. Specifically, we consider the Relational Network~\parencite{santoro2017simple} architecture as a natural generalization of the DeepSets~\parencite{zaheer2017deep} architecture, and study their representational gap. Under the restriction to analytic activation functions, we construct a symmetric function acting on sets of size \$N\$ with elements in dimension \$D\$, which can be efficiently approximated by the former architecture, but provably requires width exponential in \$N\$ and \$D\$ for the latter.

## [On Non-Linear operators for Geometric Deep Learning](#)

- Grégoire Sergeant-Perthuis · Jakob Maier · Joan Bruna · Edouard Oyallon
- abstract@[open-review](#): This work studies operators mapping vector and scalar fields defined over a manifold \$\mathcal{M}\$, and which commute with its group of diffeomorphisms \$\text{Diff}(\mathcal{M})\$. We prove that in the case of scalar fields \$L^p\omega(\mathcal{M},\mathbb{R})\$, those operators correspond to point-wise non-linearities, recovering and extending known results on \$\mathbb{R}^d\$. In the context of Neural Networks defined over \$\mathcal{M}\$, it indicates that point-wise non-linear operators are the only universal family that commutes with any group of symmetries, and justifies their systematic use in combination with dedicated linear operators commuting with specific symmetries. In the case of vector fields \$L^p\omega(\mathcal{M},T\mathcal{M})\$, we show that those operators are solely the scalar multiplication. It indicates that \$\text{Diff}(\mathcal{M})\$ is too rich and that there is no universal class of non-linear operators to motivate the design of Neural Networks over the symmetries of \$\mathcal{M}\$.

## [GlanceNets: Interpretable, Leak-proof Concept-based Models](#)

- Emanuele Marconato · Andrea Passerini · Stefano Teso
- abstract@[open-review](#): There is growing interest in concept-based models (CBMs) that combine high-performance and interpretability by acquiring and reasoning with a vocabulary of high-level concepts. A key requirement is that the concepts be interpretable. Existing CBMs tackle this desideratum using a variety of heuristics based on unclear notions of interpretability, and fail to acquire concepts with the intended semantics. We address this by providing a clear definition of interpretability in terms of alignment between the model's representation and an underlying data generation process, and introduce GlanceNets, a new CBM that exploits techniques from disentangled representation learning and open-set recognition to achieve alignment, thus improving the interpretability of the learned concepts. We show that GlanceNets, paired with concept-level supervision, achieve better alignment than state-of-the-art approaches while preventing spurious information from unintendedly leaking into the learned concepts.

## [Imitating Past Successes can be Very Suboptimal](#)

- Benjamin Eysenbach · Soumith Udatha · Russ Salakhutdinov · Sergey Levine
- abstract@[open-review](#): Prior work has proposed a simple strategy for reinforcement learning (RL): label experience with the outcomes achieved in that experience, and then imitate the relabeled experience. These outcome-conditioned imitation learning methods are appealing because of their simplicity, strong performance, and close ties with supervised learning. However, it remains unclear how these methods relate to the standard RL objective, reward maximization. In this paper, we prove that existing outcome-conditioned imitation learning methods do not necessarily improve the policy. However, we show that a simple modification results in a method that does guarantee policy improvement. Our aim is not to develop an entirely new method, but rather to explain how a variant of outcome-conditioned imitation learning can be used to maximize rewards

## [On the difficulty of learning chaotic dynamics with RNNs](#)

- Jonas Mikhaeil · Zahra Monfared · Daniel Durstewitz

- abstract@[open-review](#): Recurrent neural networks (RNNs) are wide-spread machine learning tools for modeling sequential and time series data. They are notoriously hard to train because their loss gradients backpropagated in time tend to saturate or diverge during training. This is known as the exploding and vanishing gradient problem. Previous solutions to this issue either built on rather complicated, purpose-engineered architectures with gated memory buffers, or - more recently - imposed constraints that ensure convergence to a fixed point or restrict (the eigenspectrum of) the recurrence matrix. Such constraints, however, convey severe limitations on the expressivity of the RNN. Essential intrinsic dynamics such as multistability or chaos are disabled. This is inherently at disaccord with the chaotic nature of many, if not most, time series encountered in nature and society. It is particularly problematic in scientific applications where one aims to reconstruct the underlying dynamical system. Here we offer a comprehensive theoretical treatment of this problem by relating the loss gradients during RNN training to the Lyapunov spectrum of RNN-generated orbits. We mathematically prove that RNNs producing stable equilibrium or cyclic behavior have bounded gradients, whereas the gradients of RNNs with chaotic dynamics always diverge. Based on these analyses and insights we suggest ways of how to optimize the training process on chaotic data according to the system's Lyapunov spectrum, regardless of the employed RNN architecture.

## [Differentially Private Covariance Revisited](#)

- Wei Dong Â· Yuting Liang Â· Ke Yi
- abstract@[open-review](#): In this paper, we present two new algorithms for covariance estimation under concentrated differential privacy (zCDP). The first algorithm achieves a Frobenius error of  $\tilde{O}(d^{1/4}\sqrt{\text{tr}}\sqrt{n} + \sqrt{d}/n)$ , where  $\text{tr}$  is the trace of the covariance matrix. By taking  $\text{tr}=1$ , this also implies a worst-case error bound of  $\tilde{O}(d^{1/4}\sqrt{n})$ , which improves the standard Gaussian mechanism's  $\tilde{O}(d/n)$  for the regime  $d > \tilde{\Omega}(n^{2/3})$ . Our second algorithm offers a tail-sensitive bound that could be much better on skewed data. The corresponding algorithms are also simple and efficient. Experimental results show that they offer significant improvements over prior work.

## [Why Robust Generalization in Deep Learning is Difficult: Perspective of Expressive Power](#)

- Binghui Li Â· Jikai Jin Â· Han Zhong Â· John Hopcroft Â· Liwei Wang
- abstract@[open-review](#): It is well-known that modern neural networks are vulnerable to adversarial examples. To mitigate this problem, a series of robust learning algorithms have been proposed. However, although the robust training error can be near zero via some methods, all existing algorithms lead to a high robust generalization error. In this paper, we provide a theoretical understanding of this puzzling phenomenon from the perspective of expressive power for deep neural networks. Specifically, for binary classification problems with well-separated data, we show that, for ReLU networks, while mild over-parameterization is sufficient for high robust training accuracy, there exists a constant robust generalization gap unless the size of the neural network is exponential in the data dimension  $d$ . Even if the data is linear separable, which means achieving low clean generalization error is easy, we can still prove an  $\exp(\Omega(d))$  lower bound for robust generalization. Moreover, we establish an improved upper bound of  $\exp(\mathcal{O}(k))$  for the network size to achieve low robust generalization error when the data lies on a manifold with intrinsic dimension  $k$  ( $k \leq d$ ). Nonetheless, we also have a lower bound that grows exponentially with respect to  $k$  --- the curse of dimensionality is inevitable. By demonstrating an exponential separation between the network size for achieving low robust training and generalization error, our results reveal that the hardness of robust generalization may stem from the expressive power of practical models.

## [Beyond Rewards: a Hierarchical Perspective on Offline Multiagent Behavioral Analysis](#)

- Shayegan Omidshafiei Â· Andrei Kapishnikov Â· Yannick Assogba Â· Lucas Dixon Â· Been Kim
- abstract@[open-review](#): Each year, expert-level performance is attained in increasingly-complex multiagent domains, notable examples including Go, Poker, and StarCraft II. This rapid progression is accompanied by a commensurate need to better understand how such agents attain this performance, to enable their safe deployment, identify limitations, and reveal potential means of improving them. In this paper we take a step back from performance-focused multiagent learning, and instead turn our attention towards agent behavior analysis. We introduce a model-agnostic method for discovery of behavior clusters in multiagent domains, using variational inference to learn a hierarchy of behaviors at the joint and local agent levels. Our framework makes no assumption about agents' underlying learning algorithms, does not require access to their latent states or policies, and is trained using only offline observational data. We illustrate the effectiveness of our method for enabling the coupled understanding of behaviors at the joint and local agent level, detection of behavior changepoints throughout training, discovery of core behavioral concepts, demonstrate the approach's scalability to a high-dimensional multiagent MuJoCo control domain, and also illustrate that the approach can disentangle previously-trained policies in OpenAI's hide-and-seek domain.

## [ElasticMVS: Learning elastic part representation for self-supervised multi-view stereopsis](#)

- Jinzhi Zhang Â· Ruofan Tang Â· Zheng Cao Â· Jing Xiao Â· Ruqi Huang Â· LU FANG
- abstract@[open-review](#): Self-supervised multi-view stereopsis (MVS) attracts increasing attention for learning dense surface predictions from only a set of images without onerous ground-truth 3D training data for supervision. However, existing methods highly rely on the local photometric consistency, which fails to identify accurately dense correspondence in broad textureless and reflectance areas. In this paper, we show that geometric proximity such as surface connectedness and occlusion boundaries implicitly inferred from images could serve as reliable guidance for pixel-wise multi-view correspondences. With this insight, we present a novel elastic part representation which encodes physically-connected part segmentations with elastically-varying scales, shapes and boundaries. Meanwhile, a self-supervised MVS framework namely ElasticMVS is proposed to learn the representation and estimate per-view depth following a part-aware propagation and evaluation scheme. Specifically, the pixel-wise part representation is trained by a contrastive learning-based strategy, which increases the representation compactness in geometrically concentrated areas and contrasts otherwise. ElasticMVS iteratively optimizes a part-level consistency loss and a surface smoothness loss, based on a set of depth hypotheses propagated from the geometrically concentrated parts. Extensive evaluations convey the superiority of ElasticMVS in the reconstruction completeness and accuracy, as well as the efficiency and scalability. Particularly, for the challenging large-scale reconstruction benchmark, ElasticMVS demonstrates significant performance gain over both the supervised and self-supervised approaches.

## [Second Thoughts are Best: Learning to Re-Align With Human Values from Text Edits](#)

- Ruibo Liu Â· Chenyan Jia Â· Ge Zhang Â· Ziyu Zhuang Â· Tony Liu Â· Soroush Vosoughi
- abstract@[open-review](#): We present Second Thoughts, a new learning paradigm that enables language models (LMs) to re-align with human values. By modeling the chain-of-edits between value-unaligned and value-aligned text, with LM fine-tuning and additional refinement through reinforcement learning, Second Thoughts not only achieves superior performance in three value alignment benchmark datasets but also shows strong human-value transfer learning ability in few-shot scenarios. The generated editing steps also offer better interpretability and ease for interactive error correction. Extensive human evaluations further confirm its effectiveness.

## [LST: Ladder Side-Tuning for Parameter and Memory Efficient Transfer Learning](#)

- Yi-Lin Sung Â· Jaemin Cho Â· Mohit Bansal
- abstract@[open-review](#): Fine-tuning large pre-trained models on downstream tasks has been adopted in a variety of domains recently. However, it is costly to update the entire parameter set of large pre-trained models. Although recently proposed parameter-efficient transfer learning (PETL) techniques allow updating a small subset of parameters inside a pre-trained backbone network for a new task, they only reduce the training memory requirement by up to

\$30\%\$. This is because the gradient computation for the trainable parameters still requires backpropagation through the large pre-trained backbone model. To address this, we propose Ladder Side-Tuning (LST), a new PETL technique that can also reduce training memory requirements by more substantial amounts. Unlike existing parameter-efficient methods that insert additional parameters inside backbone networks, we train a ladder side network, a small and separate network that takes intermediate activations as input via shortcut connections (ladders) from backbone networks and makes predictions. LST has significantly lower memory requirements than previous methods, because it does not require backpropagation through the backbone network, but instead only through the side network and ladder connections. We conduct our experiments on various models (T5 and CLIP-T5) and tasks (GLUE, VQA, GQA, NLVR\$^2\$, MSCOCO). LST saves \$69\%\$ of the memory costs to fine-tune the whole network, while other methods only save \$26\%\$ of that in similar parameter usages (hence, 2.7x more memory savings). Moreover, LST outperforms Adapter and LoRA in a low-memory regime. To further show the advantage of this memory efficiency, we also apply LST to large T5 models (T5-large, T5-3B), attaining better GLUE performance than full fine-tuning and other PETL methods. The trend also holds in the experiments on vision-and-language tasks, where LST performs comparably to other PETL methods when training a similar number of parameters but only uses \$46\%\$ of their GPU memory.

## [Beyond Time-Average Convergence: Near-Optimal Uncoupled Online Learning via Clairvoyant Multiplicative Weights Update](#)

- Georgios Piliouras · Ryann Sim · EFSTRATIOS SKOULAKIS
- abstract@[open-review](#): In this paper, we provide a novel and simple algorithm, Clairvoyant Multiplicative Weights Updates (CMWU) for regret minimization in general games. CMWU effectively corresponds to the standard MWU algorithm but where all agents, when updating their mixed strategies, use the payoff profiles based on tomorrow's behavior, i.e. the agents are clairvoyant. CMWU achieves constant regret of  $\ln(m)\eta$  in all normal-form games with  $m$  actions and fixed step-sizes  $\eta$ . Although CMWU encodes in its definition a fixed point computation, which in principle could result in dynamics that are neither computationally efficient nor uncoupled, we show that both of these issues can be largely circumvented. Specifically, as long as the step-size  $\eta$  is upper bounded by  $\frac{1}{(n-1)V}$ , where  $n$  is the number of agents and  $[0, V]$  is the payoff range, then the CMWU updates can be computed linearly fast via a contraction map. This implementation results in an uncoupled online learning dynamic that admits a  $O(\log T)$ -sparse subsequence where each agent experiences at most  $O(nV \log m)$  regret. This implies that the CMWU dynamics converge with rate  $O(nV \log m \log T / T)$  to a Coarse Correlated Equilibrium}.

## [Drawing out of Distribution with Neuro-Symbolic Generative Models](#)

- Yichao Liang · Josh Tenenbaum · Tuan Anh Le · Siddharth N
- abstract@[open-review](#): Learning general-purpose representations from perceptual inputs is a hallmark of human intelligence. For example, people can write out numbers or characters, or even draw doodles, by characterizing these tasks as different instantiations of the same generic underlying process---compositional arrangements of different forms of pen strokes. Crucially, learning to do one task, say writing, implies reasonable competence at another, say drawing, on account of this shared process. We present Drawing out of Distribution (DooD), a neuro-symbolic generative model of stroke-based drawing that can learn such general-purpose representations. In contrast to prior work, DooD operates directly on images, requires no supervision or expensive test-time inference, and performs unsupervised amortized inference with a symbolic stroke model that better enables both interpretability and generalization. We evaluate DooD on its ability to generalize across both data and tasks. We first perform zero-shot transfer from one dataset (e.g. MNIST) to another (e.g. Quickdraw), across five different datasets, and show that DooD clearly outperforms different baselines. An analysis of the learnt representations further highlights the benefits of adopting a symbolic stroke model. We then adopt a subset of the Omniglot challenge tasks, and evaluate its ability to generate new exemplars (both unconditionally and conditionally), and perform one-shot classification, showing that DooD matches the state of the art. Taken together, we demonstrate that DooD does indeed capture general-purpose representations across both data and task, and takes a further step towards building general and robust concept-learning systems.

## [Rashomon Capacity: A Metric for Predictive Multiplicity in Probabilistic Classification](#)

- Hsiang Hsu · Flavio Calmon
- abstract@[open-review](#): Predictive multiplicity occurs when classification models with nearly indistinguishable average performances assign conflicting predictions to individual samples. When used for decision-making in applications of consequence (e.g., lending, education, criminal justice), models developed without regard for predictive multiplicity may result in unjustified and arbitrary decisions for specific individuals. We introduce a new measure of predictive multiplicity in probabilistic classification called Rashomon capacity. Prior metrics for predictive multiplicity focus on classifiers that output thresholded (i.e., 0-1) predicted classes. In contrast, Rashomon capacity applies to probabilistic classifiers, capturing more nuanced score variations for individual samples. We provide a rigorous derivation for Rashomon capacity, argue its intuitive appeal, and demonstrate how to estimate it in practice. We show that Rashomon capacity yields principled strategies for disclosing conflicting models to stakeholders. Our numerical experiments illustrate how Rashomon capacity captures predictive multiplicity in various datasets and learning models, including neural networks. The tools introduced in this paper can help data scientists measure, report, and ultimately resolve predictive multiplicity prior to model deployment.

## [Cache-Augmented Inbatch Importance Resampling for Training Recommender Retriever](#)

- Jin Chen · Defu Lian · Yucheng Li · Baoyun Wang · Kai Zheng · Enhong Chen
- abstract@[open-review](#): Recommender retrievers aim to rapidly retrieve a fraction of items from the entire item corpus when a user query requests, with the representative two-tower model trained with the log softmax loss. For efficiently training recommender retrievers on modern hardwares, inbatch sampling, where the items in the mini-batch are shared as negatives to estimate the softmax function, has attained growing interest. However, existing inbatch sampling based strategies just correct the sampling bias of inbatch items with item frequency, being unable to distinguish the user queries within the mini-batch and still incurring significant bias from the softmax. In this paper, we propose a Cache-Augmented Inbatch Importance Resampling (XIR) for training recommender retrievers, which not only offers different negatives to user queries with inbatch items, but also adaptively achieves a more accurate estimation of the softmax distribution. Specifically, XIR resamples items for the given mini-batch training pairs based on certain probabilities, where a cache with more frequently sampled items is adopted to augment the candidate item set, with the purpose of reusing the historical informative samples. XIR enables to sample query-dependent negatives based on inbatch items and to capture dynamic changes of model training, which leads to a better approximation of the softmax and further contributes to better convergence. Finally, we conduct experiments to validate the superior performance of the proposed XIR compared with competitive approaches.

## [Towards Robust Blind Face Restoration with Codebook Lookup Transformer](#)

- Shangchen Zhou · Kelvin Chan · Chongyi Li · Chen Change Loy
- abstract@[open-review](#): Blind face restoration is a highly ill-posed problem that often requires explicit auxiliary guidance to improve the mapping from a degraded input to a desired restored output. In this paper, we demonstrate that the uncertainty and ambiguity of the mapping can be largely reduced by casting face restoration as a code prediction task in a small, finite proxy feature space. Under this paradigm, we propose a Transformer-based prediction network, named \textit{CodeFormer}, to exploit global contexts of the input for \textit{code prediction}, enabling the discovery of a natural face that closely approximates the target high-quality image even when the input is severely degraded. To further enhance identity preservation, we propose a controllable feature transformation module to control information flow from the input image. Such a design allows a flexible trade-off between fidelity and quality so that one could reduce the reliance on the input image in case of heavy degradation. Equipped with the proposed components, our \textit{CodeFormer} suggests superior robustness against degradation, outperforming state of the arts in both quality and fidelity. Extensive analysis is also conducted to verify the effectiveness of our method.

## [Fast Mixing of Stochastic Gradient Descent with Normalization and Weight Decay](#)

- Zhiyuan Li · Tianhao Wang · Dingli Yu
- abstract@[open-review](#): We prove the Fast Equilibrium Conjecture [26], i.e., Stochastic Gradient Descent (SGD) for scale-invariant loss (e.g., networks with various normalization schemes) with learning rate  $\eta$  and weight decay factor  $\lambda$  mixes in function space in  $\mathcal{O}$  ( $\frac{1}{\lambda\eta}$ ) steps, under two standard assumptions: (1) the noise covariance matrix is non-degenerate and has constant trace and (2) all minimizers of the loss form a connected and smooth manifold. The analysis uses the framework from [27] and shows that for every  $T > 0$ , the iterates of SGD with learning rate  $\eta$  and weight decay factor  $\lambda$  on the scale-invariant loss converge in distribution in  $\Theta(\left(\eta^{-1}\lambda^{-1}(T + \ln(\lambda/\eta))\right)^{-1})$  iterations as  $\eta, \lambda \rightarrow 0$  (while simultaneously satisfying  $\eta \leq \lambda$  and  $\lambda \leq 1$ ). Moreover, the evolution of limiting distribution can be described by an Ito's stochastic differential equation (SDE) which mixes to a unique distribution in time independent of  $\eta$ .

## [Fully Sparse 3D Object Detection](#)

- Lue Fan · Feng Wang · Naiyan Wang · ZHAO-XIANG ZHANG
- abstract@[open-review](#): As the perception range of LiDAR increases, LiDAR-based 3D object detection becomes a dominant task in the long-range perception task of autonomous driving. The mainstream 3D object detectors usually build dense feature maps in the network backbone and prediction head. However, the computational and spatial costs on the dense feature map are quadratic to the perception range, which makes them hardly scale up to the long-range setting. To enable efficient long-range LiDAR-based object detection, we build a fully sparse 3D object detector (FSD). The computational and spatial cost of FSD is roughly linear to the number of points and independent of the perception range. FSD is built upon the general sparse voxel encoder and a novel sparse instance recognition (SIR) module. SIR first groups the points into instances and then applies instance-wise feature extraction and prediction. In this way, SIR resolves the issue of center feature missing, which hinders the design of the fully sparse architecture for all center-based or anchor-based detectors. Moreover, SIR avoids the time-consuming neighbor queries in previous point-based methods by grouping points into instances. We conduct extensive experiments on the large-scale Waymo Open Dataset to reveal the working mechanism of FSD, and state-of-the-art performance is reported. To demonstrate the superiority of FSD in long-range detection, we also conduct experiments on Argoverse 2 Dataset, which has a much larger perception range (\$200m\$) than Waymo Open Dataset (\$75m\$). On such a large perception range, FSD achieves state-of-the-art performance and is 2.4\$ times faster than the dense counterpart. Codes will be released.

## [VICRegL: Self-Supervised Learning of Local Visual Features](#)

- Adrien Bardes · Jean Ponce · Yann LeCun
- abstract@[open-review](#): Most recent self-supervised methods for learning image representations focus on either producing a global feature with invariance properties, or producing a set of local features. The former works best for classification tasks while the latter is best for detection and segmentation tasks. This paper explores the fundamental trade-off between learning local and global features. A new method called VICRegL is proposed that learns good global and local features simultaneously, yielding excellent performance on detection and segmentation tasks while maintaining good performance on classification tasks. Concretely, two identical branches of a standard convolutional net architecture are fed two differently distorted versions of the same image. The VICReg criterion is applied to pairs of global feature vectors. Simultaneously, the VICReg criterion is applied to pairs of local feature vectors occurring before the last pooling layer. Two local feature vectors are attracted to each other if their l2-distance is below a threshold or if their relative locations are consistent with a known geometric transformation between the two input images. We demonstrate strong performance on linear classification and segmentation transfer tasks.

## [Advancing Model Pruning via Bi-level Optimization](#)

- Yihua Zhang · Yuguang Yao · Parikshit Ram · pu zhao · Tianlong Chen · Mingyi Hong · Yanzhi Wang · Sijia Liu
- abstract@[open-review](#): The deployment constraints in practical applications necessitate the pruning of large-scale deep learning models, i.e., promoting their weight sparsity. As illustrated by the Lottery Ticket Hypothesis (LTH), pruning also has the potential of improving their generalization ability. At the core of LTH, iterative magnitude pruning (IMP) is the predominant pruning method to successfully find “winning tickets”. Yet, the computation cost of IMP grows prohibitively as the targeted pruning ratio increases. To reduce the computation overhead, various efficient “one-shot” pruning methods have been developed, but these schemes are usually unable to find winning tickets as good as IMP. This raises the question of how to close the gap between pruning accuracy and pruning efficiency? To tackle it, we pursue the algorithmic advancement of model pruning. Specifically, we formulate the pruning problem from a fresh and novel viewpoint, bi-level optimization (BLO). We show that the BLO interpretation provides a technically-grounded optimization base for an efficient implementation of the pruning-retraining learning paradigm used in IMP. We also show that the proposed bi-level optimization-oriented pruning method (termed BiP) is a special class of BLO problems with a bi-linear problem structure. By leveraging such bi-linearity, we theoretically show that BiP can be solved as easily as first-order optimization, thus inheriting the computation efficiency. Through extensive experiments on both structured and unstructured pruning with 5 model architectures and 4 data sets, we demonstrate that BiP can find better winning tickets than IMP in most cases, and is computationally as efficient as the one-shot pruning schemes, demonstrating 2-7\$ times\$ speedup over IMP for the same level of model accuracy and sparsity.

## [Learning With an Evolving Class Ontology](#)

- Zhiqiu Lin · Yu-Xiong Wang · Deepak Pathak · Deva Ramanan · Shu Kong
- abstract@[open-review](#): Lifelong learners must recognize concept vocabularies that evolve over time. A common yet underexplored scenario is learning with class labels over time that refine/expand old classes. For example, humans learn to recognize “dog” before dog breeds. In practical settings, dataset “versioning” often introduces refinement to ontologies, such as autonomous vehicle benchmarks that refine a previous “vehicle” class into “school-bus” as autonomous operations expand to new cities. This paper formalizes a protocol for studying the problem of “Learning with Evolving Class Ontology” (LECO). LECO requires learning classifiers in distinct time periods (TPs); each TP introduces a new ontology of “fine” labels that refines old ontologies of “coarse” labels (e.g., dog breeds that refine the previous “dog”). LECO explores such questions as whether to annotate new data or relabel the old, how to leverage coarse labels, and whether to finetune the previous TP’s model or train from scratch. To answer these questions, we leverage insights from related problems such as class-incremental learning. We validate them under the LECO protocol through the lens of image classification (on CIFAR and iNaturalist) and semantic segmentation (on Mapillary). Extensive experiments lead to some surprising conclusions; while the current status quo in the field is to relabel existing datasets with new class ontologies (such as COCO-to-LVIS or Mapillary1.2-to-2.0), LECO demonstrates that a far better strategy is to annotate “new” data with the new ontology. However, this produces an aggregate dataset with inconsistent old-vs-new labels, complicating learning. To address this challenge, we adopt methods from semi-supervised and partial-label learning. We demonstrate that such strategies can surprisingly be made near-optimal, in the sense of approaching an “oracle” that learns on the aggregate dataset exhaustively labeled with the newest ontology.

## [Optimal Gradient Sliding and its Application to Optimal Distributed Optimization Under Similarity](#)

- Dmitry Kovalev · Aleksandr Beznosikov · Ekaterina Borodich · Alexander Gasnikov · Gesualdo Scutari
- abstract@[open-review](#): We study structured convex optimization problems, with additive objective  $r := p + q$ , where  $r$  is ( $\mu$ -strongly) convex,  $q$  is  $L_q$ -smooth and convex, and  $p$  is  $L_p$ -smooth, possibly nonconvex. For such a class of problems, we proposed an inexact accelerated gradient sliding method that can skip the gradient computation for one of these components while still achieving optimal complexity of gradient calls of  $p$  and

$\$q$ , that is,  $\$mathcal{O}(\sqrt{L_p/\mu})$  and  $\$mathcal{O}(\sqrt{L_q/\mu})$ , respectively. This result is much sharper than the classic black-box complexity  $\$mathcal{O}(\sqrt{(L_p+L_q)/\mu})$ , especially when the difference between  $L_p$  and  $L_q$  is large. We then apply the proposed method to solve distributed optimization problems over master-worker architectures, under agents' function similarity, due to statistical data similarity or otherwise. The distributed algorithm achieves for the first time lower complexity bounds on {\it if both} communication and local gradient calls, with the former having being a long-standing open problem. Finally the method is extended to distributed saddle-problems (under function similarity) by means of solving a class of variational inequalities, achieving lower communication and computation complexity bounds.

## [Semi-Parametric Neural Image Synthesis](#)

- Andreas Blattmann · Robin Rombach · Kaan Oktay · Jonas Mäller · Björn Ommer
- abstract@[open-review](#): Novel architectures have recently improved generative image synthesis leading to excellent visual quality in various tasks. Much of this success is due to the scalability of these architectures and hence caused by a dramatic increase in model complexity and in the computational resources invested in training these models. Our work questions the underlying paradigm of compressing large training data into ever growing parametric representations. We rather present an orthogonal, semi-parametric approach. We complement a comparably small generative image model, e.g., a diffusion or autoregressive model, with a separate image database and a retrieval-based approach. During training we retrieve a set of nearest neighbors from this external database for each training instance and condition the generative model on these informative samples. While the retrieval approach is providing the (local) content, the model is focusing on learning the composition of scenes based on this content. As demonstrated by our experiments, simply swapping the database for one with different contents transfers a trained model post-hoc to a novel domain. The evaluation shows competitive performance on tasks which the generative model has not been trained on, such as class-conditional or text-to-image synthesis and zero-shot stylization. With negligible memory and computational overhead for external database and retrieval we can significantly reduce the parameter count of the generative model and still outperform the state-of-the-art.

## [Masked Autoencoders As Spatiotemporal Learners](#)

- Christoph Feichtenhofer · Haoqi Fan · Yanghao Li · Kaiming He
- abstract@[open-review](#): This paper studies a conceptually simple extension of Masked Autoencoders (MAE) to spatiotemporal representation learning from videos. We randomly mask out spacetime patches in videos and learn an autoencoder to reconstruct them in pixels. Interestingly, we show that our MAE method can learn strong representations with almost no inductive bias on spacetime (only except for patch and positional embeddings), and spacetime-agnostic random masking performs the best. We observe that the optimal masking ratio is as high as 90% (vs. 75% on images), supporting the hypothesis that this ratio is related to information redundancy of the data. A high masking ratio leads to a large speedup, e.g., > 4x in wall-clock time or even more. We report competitive results on several challenging video datasets using vanilla Vision Transformers. We observe that MAE can outperform supervised pre-training by large margins. We further report encouraging results of training on real-world, uncurated Instagram data. Our study suggests that the general framework of masked autoencoding (BERT, MAE, etc.) can be a unified methodology for representation learning with minimal domain knowledge.

## [ELASTIC: Numerical Reasoning with Adaptive Symbolic Compiler](#)

- Jiaxin Zhang · Yashar Moshfeghi
- abstract@[open-review](#): Numerical reasoning over text is a challenging task of Artificial Intelligence (AI), requiring reading comprehension and numerical reasoning abilities. Previous approaches use numerical reasoning programs to represent the reasoning process. However, most works do not separate the generation of operators and operands, which are key components of a numerical reasoning program, thus limiting their ability to generate such programs for complicated tasks. In this paper, we introduce the numEricaL reASoning with adapTive symbolIc Compiler (ELASTIC) model, which is constituted of the RoBERTa as the Encoder and a Compiler with four modules: Reasoning Manager, Operator Generator, Operands Generator, and Memory Register. ELASTIC is robust when conducting complicated reasoning. Also, it is domain agnostic by supporting the expansion of diverse operators without caring about the number of operands it contains. Experiments show that ELASTIC achieves 68.96 and 65.21 of execution accuracy and program accuracy on the FinQA dataset and 83.00 program accuracy on the MathQA dataset, outperforming previous state-of-the-art models significantly.

## [Sequential Information Design: Learning to Persuade in the Dark](#)

- Martino Bernasconi · Matteo Castiglioni · Alberto Marchesi · Nicola Gatti · Francesco Trovò<sup>2</sup>
- abstract@[open-review](#): We study a repeated information design problem faced by an informed sender who tries to influence the behavior of a self-interested receiver. We consider settings where the receiver faces a sequential decision making (SDM) problem. At each round, the sender observes the realizations of random events in the SDM problem. This begets the challenge of how to incrementally disclose such information to the receiver to persuade them to follow (desirable) action recommendations. We study the case in which the sender does not know random events probabilities, and, thus, they have to gradually learn them while persuading the receiver. Our goal is to design online learning algorithms that are no-regret for the sender, while at the same time being persuasive for the receiver. We start by providing a non-trivial polytopal approximation of the set of sender's persuasive information structures. This is crucial to design efficient learning algorithms. Next, we prove a negative result: no learning algorithm can be persuasive. Thus, we relax persuasiveness requirements by focusing on algorithms that guarantee that the receiver's regret in following recommendations grows sub-linearly. In the full-feedback setting---where the sender observes all random events realizations---, we provide an algorithm with  $\tilde{O}(\sqrt{T})$  regret for both the sender and the receiver. Instead, in the bandit-feedback setting---where the sender only observes the realizations of random events actually occurring in the SDM problem---, we design an algorithm that, given an  $\alpha \in [1/2, 1]$  as input, ensures  $\tilde{O}(T^{\alpha})$  and  $\tilde{O}(T^{\max\{\alpha, 1-\frac{\alpha}{2}\}})$  regrets for the sender and the receiver, respectively. This result is complemented by a lower bound showing that such a regrets trade-off is essentially tight.

## [General Cutting Planes for Bound-Propagation-Based Neural Network Verification](#)

- Huan Zhang · Shiqi Wang · Kaidi Xu · Linyi Li · Bo Li · Suman Jana · Cho-Jui Hsieh · J. Zico Kolter
- abstract@[open-review](#): Bound propagation methods, when combined with branch and bound, are among the most effective methods to formally verify properties of deep neural networks such as correctness, robustness, and safety. However, existing works cannot handle the general form of cutting plane constraints widely accepted in traditional solvers, which are crucial for strengthening verifiers with tightened convex relaxations. In this paper, we generalize the bound propagation procedure to allow the addition of arbitrary cutting plane constraints, including those involving relaxed integer variables that do not appear in existing bound propagation formulations. Our generalized bound propagation method, GCP-CROWN, opens up the opportunity to apply general cutting plane methods for neural network verification while benefiting from the efficiency and GPU acceleration of bound propagation methods. As a case study, we investigate the use of cutting planes generated by off-the-shelf mixed integer programming (MIP) solver. We find that MIP solvers can generate high-quality cutting planes for strengthening bound-propagation-based verifiers using our new formulation. Since the branching-focused bound propagation procedure and the cutting-plane-focused MIP solver can run in parallel utilizing different types of hardware (GPUs and CPUs), their combination can quickly explore a large number of branches with strong cutting planes, leading to strong verification performance. Experiments demonstrate that our method is the first verifier that can completely solve the oval20 benchmark, and can verify twice as many instances on the oval21 benchmark compared to the best tool in VNN-COMP 2021, and also noticeably outperforms state-of-the-art verifiers on a wide range of benchmarks.

## [Oracle Inequalities for Model Selection in Offline Reinforcement Learning](#)

- Jonathan N Lee · George Tucker · Ofir Nachum · Bo Dai · Emma Brunskill
- abstract@[open-review](#): Offline reinforcement learning (RL) is a promising paradigm where a learner leverages prior data to learn a good policy without interacting with the environment. A major challenge in applying such methods in practice is the lack of both theoretically principled and practical tools for model selection and evaluation. To address this, we study the problem of model selection in offline RL with value function approximation where the learner is given a nested sequence of model classes to minimize squared Bellman error and must select among these to achieve the optimal balance of approximation and estimation error of the classes. We propose, to our knowledge, the first model selection algorithm for offline RL that achieves minimax rate-optimal oracle inequalities up to logarithmic factors. The algorithm, ModBE, takes as input the model classes and a base offline RL algorithm designed to minimize squared Bellman error. It successively eliminates model classes using a novel one-sided generalization test, finally returning a policy that competes with the performance of the best model class. In addition to its theoretical guarantees, it is conceptually simple and computationally efficient, amounting to calculating and comparing relative squared errors between classes. Finally, we demonstrate it is capable of reliably selecting a good model class in small simulated experiments.

## [Adjoint-aided inference of Gaussian process driven differential equations](#)

- Paterne GAHUNGU · Christopher Lanyon · Mauricio A Álvarez · Engineer Bainomugisha · Michael T Smith · Richard Wilkinson
- abstract@[open-review](#): Linear systems occur throughout engineering and the sciences, most notably as differential equations. In many cases the forcing function for the system is unknown, and interest lies in using noisy observations of the system to infer the forcing, as well as other unknown parameters. In differential equations, the forcing function is an unknown function of the independent variables (typically time and space), and can be modelled as a Gaussian process (GP). In this paper we show how the adjoint of a linear system can be used to efficiently infer forcing functions modelled as GPs, after using a truncated basis expansion of the GP kernel. We show how exact conjugate Bayesian inference for the truncated GP can be achieved, in many cases with substantially lower computation than would be required using MCMC methods. We demonstrate the approach on systems of both ordinary and partial differential equations, and show that the basis expansion approach approximates well the true forcing with a modest number of basis vectors. Finally, we show how to infer point estimates for the non-linear model parameters, such as the kernel length-scales, using Bayesian optimisation.

## [Uncertain Estimation for Multi-view Data: The Power of Seeing the Whole Picture](#)

- Myong Chol Jung · He Zhao · Joanna Dipnall · Belinda Gabbe · Lan Du
- abstract@[open-review](#): Uncertainty estimation is essential to make neural networks trustworthy in real-world applications. Extensive research efforts have been made to quantify and reduce predictive uncertainty. However, most existing works are designed for unimodal data, whereas multi-view uncertainty estimation has not been sufficiently investigated. Therefore, we propose a new multi-view classification framework for better uncertainty estimation and out-of-domain sample detection, where we associate each view with an uncertainty-aware classifier and combine the predictions of all the views in a principled way. The experimental results with real-world datasets demonstrate that our proposed approach is an accurate, reliable, and well-calibrated classifier, which predominantly outperforms the multi-view baselines tested in terms of expected calibration error, robustness to noise, and accuracy for the in-domain sample classification and the out-of-domain sample detection tasks.

## [Randomized Message-Interception Smoothing: Gray-box Certificates for Graph Neural Networks](#)

- Yan Scholten · Jan Schuchardt · Simon Geisler · Aleksandar Bojchevski · Stephan Günnemann
- abstract@[open-review](#): Randomized smoothing is one of the most promising frameworks for certifying the adversarial robustness of machine learning models, including Graph Neural Networks (GNNs). Yet, existing randomized smoothing certificates for GNNs are overly pessimistic since they treat the model as a black box, ignoring the underlying architecture. To remedy this, we propose novel gray-box certificates that exploit the message-passing principle of GNNs: We randomly intercept messages and carefully analyze the probability that messages from adversarially controlled nodes reach their target nodes. Compared to existing certificates, we certify robustness to much stronger adversaries that control entire nodes in the graph and can arbitrarily manipulate node features. Our certificates provide stronger guarantees for attacks at larger distances, as messages from farther-away nodes are more likely to get intercepted. We demonstrate the effectiveness of our method on various models and datasets. Since our gray-box certificates consider the underlying graph structure, we can significantly improve certifiable robustness by applying graph sparsification.

## [Task-Agnostic Graph Explanations](#)

- Yaochen Xie · Sumeet Katariya · Xianfeng Tang · Edward Huang · Nikhil Rao · Karthik Subbian · Shuiwang Ji
- abstract@[open-review](#): Graph Neural Networks (GNNs) have emerged as powerful tools to encode graph-structured data. Due to their broad applications, there is an increasing need to develop tools to explain how GNNs make decisions given graph-structured data. Existing learning-based GNN explanation approaches are task-specific in training and hence suffer from crucial drawbacks. Specifically, they are incapable of producing explanations for a multitask prediction model with a single explainer. They are also unable to provide explanations in cases where the GNN is trained in a self-supervised manner, and the resulting representations are used in future downstream tasks. To address these limitations, we propose a Task-Agnostic GNN Explainer (TAGE) that is independent of downstream models and trained under self-supervision with no knowledge of downstream tasks. TAGE enables the explanation of GNN embedding models with unseen downstream tasks and allows efficient explanation of multitask models. Our extensive experiments show that TAGE can significantly speed up the explanation efficiency by using the same model to explain predictions for multiple downstream tasks while achieving explanation quality as good as or even better than current state-of-the-art GNN explanation approaches.

## [Differentially Private Linear Sketches: Efficient Implementations and Applications](#)

- Fuheng Zhao · Dan Qiao · Rachel Redberg · Divyakant Agrawal · Amr El Abbadi · Yu-Xiang Wang
- abstract@[open-review](#): Linear sketches have been widely adopted to process fast data streams, and they can be used to accurately answer frequency estimation, approximate top K items, and summarize data distributions. When data are sensitive, it is desirable to provide privacy guarantees for linear sketches to preserve private information while delivering useful results with theoretical bounds. We show that linear sketches can ensure privacy and maintain their unique properties with a small amount of noise added at initialization. From the differentially private linear sketches, we showcase that the state-of-the-art quantile sketch in the turnstile model can also be private and maintain high performance. Experiments further demonstrate that our proposed differentially private sketches are quantitatively and qualitatively similar to noise-free sketches with high utilization on synthetic and real datasets.

## [Learning Viewpoint-Agnostic Visual Representations by Recovering Tokens in 3D Space](#)

- Jinghuan Shang · Sriyan Das · Michael Ryoo
- abstract@[open-review](#): Humans are remarkably flexible in understanding viewpoint changes due to visual cortex supporting the perception of 3D structure. In contrast, most of the computer vision models that learn visual representation from a pool of 2D images often fail to generalize over novel camera viewpoints. Recently, the vision architectures have shifted towards convolution-free architectures, visual Transformers, which operate on tokens derived from image patches. However, these Transformers do not perform explicit operations to learn viewpoint-agnostic representation for visual understanding, as in convolutions. To this end, we propose a 3D Token Representation Layer (3DTRL) that estimates the 3D positional information of the visual tokens and leverages it for learning viewpoint-agnostic representations. The key elements of 3DTRL include a pseudo-depth estimator and a learned camera matrix to impose geometric transformations on the tokens. These enable 3DTRL to recover the 3D positional information of the tokens from 2D patches. In practice, 3DTRL is easily plugged-in into a Transformer. Our experiments demonstrate the effectiveness of 3DTRL in many vision

tasks including image classification, multi-view video alignment, and action recognition. The models with 3DTRL outperform their backbone Transformers in all the tasks with minimal added computation.

## [Robustness to Unbounded Smoothness of Generalized SignSGD](#)

- Michael Crawshaw · Mingrui Liu · Francesco Orabona · Wei Zhang · Zhenxun Zhuang
- abstract@[open-review](#): Traditional analyses in non-convex optimization typically rely on the smoothness assumption, namely requiring the gradients to be Lipschitz. However, recent evidence shows that this smoothness condition does not capture the properties of some deep learning objective functions, including the ones involving Recurrent Neural Networks and LSTMs. Instead, they satisfy a much more relaxed condition, with potentially unbounded smoothness. Under this relaxed assumption, it has been theoretically and empirically shown that the gradient-clipped SGD has an advantage over the vanilla one. In this paper, we show that clipping is not indispensable for Adam-type algorithms in tackling such scenarios: we theoretically prove that a generalized SignSGD algorithm can obtain similar convergence rates as SGD with clipping but does not need explicit clipping at all. This family of algorithms on one end recovers SignSGD and on the other end closely resembles the popular Adam algorithm. Our analysis underlines the critical role that momentum plays in analyzing SignSGD-type and Adam-type algorithms: it not only reduces the effects of noise, thus removing the need for large mini-batch in previous analyses of SignSGD-type algorithms, but it also substantially reduces the effects of unbounded smoothness and gradient norms. We also compare these algorithms with popular optimizers on a set of deep learning tasks, observing that we can match the performance of Adam while beating the others.

## [Sparse Fourier Backpropagation in Cryo-EM Reconstruction](#)

- Dari Kimanis · Kiarash Jamali · Sjors Scheres
- abstract@[open-review](#): Cryogenic electron microscopy (cryo-EM) is a powerful method for investigating the structures of protein molecules, with important implications for understanding the molecular processes of life and drug development. In this technique, many noisy, two-dimensional projection images of protein molecules in unknown poses are combined into one or more three-dimensional reconstructions. The presence of multiple structural states in the data represents a major bottleneck in existing processing pipelines, often requiring expert user supervision. Variational auto-encoders (VAEs) have recently been proposed as an attractive means of learning the data manifold for data sets with a large number of different states. These methods are based on a coordinate-based approach, similar to Neural Radiance Fields (NeRF), to make volumetric reconstructions from 2D image data in Fourier space. Although NeRF is a powerful method for real-space reconstruction, many of the benefits of the method do not transfer to Fourier space, e.g. inductive bias for spacial locality. We present an approach where the VAE reconstruction is expressed on a volumetric grid, and demonstrate how such a model can be trained efficiently through a novel backpropagation protocol that exploits the sparsity of the projection operation in Fourier space. Comparing to the coordinate-based approach, we achieve improved results for an experimental dataset with multiple structural states. Moreover, our approach is computationally more efficient, especially in inference, enabling interactive analysis of latent space by the user.

## [Alignment as a Multi-agent Intrinsic Reward](#)

- Zixian Ma · Rose Wang · Michael Bernstein · Fei-Fei Li · Ranjay Krishna
- abstract@[open-review](#): Modern multi-agent reinforcement learning frameworks rely on centralized training and reward shaping to perform well. However, centralized training and dense rewards are not readily available in the real world. Current multi-agent algorithms struggle to learn in the alternative setup of decentralized training or sparse rewards. To address these issues, we propose a self-supervised intrinsic reward called \textit{alignment} inspired by the self-organization principle in Zoology. Similar to how animals collaborate in a decentralized manner with those in their vicinity, agents trained with alignment learn behaviors that match their neighbors' expectations. This allows the agents to learn collaborative behaviors without any external reward or centralized training. We demonstrate the efficacy of our approach across 6 tasks in the multi-agent particle and the complex Google Research football environments, comparing \textit{alignment} to sparse and curiosity-based intrinsic rewards. When the number of agents increases, alignment scales well in all multi-agent tasks except for one where agents have different capabilities. We show that agent coordination improves through alignment because agents learn to divide tasks amongst themselves, break coordination symmetries, and confuse adversaries. These results identify tasks where alignment is a more useful strategy than curiosity-driven exploration for multi-agent coordination, enabling agents to do zero-shot coordination.

## [OmniVL: One Foundation Model for Image-Language and Video-Language Tasks](#)

- Junke Wang · Dongdong Chen · Zuxuan Wu · Chong Luo · Luwei Zhou · Yucheng Zhao · Yujia Xie · Ce Liu · Yu-Gang Jiang · Lu Yuan
- abstract@[open-review](#): This paper presents OmniVL, a new foundation model to support both image-language and video-language tasks using one universal architecture. It adopts a unified transformer-based visual encoder for both image and video inputs, and thus can perform joint image-language and video-language pretraining. We demonstrate, for the first time, such a paradigm benefits both image and video tasks, as opposed to the conventional one-directional transfer (e.g., use image-language to help video-language). To this end, we propose a \emph{decoupled} joint pretraining of image-language and video-language to effectively decompose the vision-language modeling into spatial and temporal dimensions and obtain performance boost on both image and video tasks. Moreover, we introduce a novel unified vision-language contrastive (UniVLC) loss to leverage image-text, video-text, image-label (e.g., image classification), video-label (e.g., video action recognition) data together, so that both supervised and noisily supervised pretraining data are utilized as much as possible. Without incurring extra task-specific adaptors, OmniVL can simultaneously support visual only tasks (e.g., image classification, video action recognition), cross-modal alignment tasks (e.g., image/video-text retrieval), and multi-modal understanding and generation tasks (e.g., image/video question answering, captioning). We evaluate OmniVL on a wide range of downstream tasks and achieve state-of-the-art or competitive results with similar model size and data scale.

## [Back Razor: Memory-Efficient Transfer Learning by Self-Sparsified Backpropagation](#)

- Ziyu Jiang · Xuxi Chen · Xueqin Huang · Xianzhi Du · Denny Zhou · Zhangyang Wang
- abstract@[open-review](#): Transfer learning from the model trained on large datasets to customized downstream tasks has been the de-facto choice as the pre-trained model can greatly boost the generalizability. However, the increasing sizes of pre-trained models also lead to a prohibitively large memory footprints for downstream transferring, making them unaffordable for personal devices. Previous work recognizes the bottleneck of the footprint to be the activation, and hence proposes various solutions such as injecting specific lite modules. In this work, we present a novel and general framework called Back Razor, that can be plug-and-play applied to any pre-trained network without changing its architecture. The key idea of Back Razor is asymmetric sparsifying: pruning the activation stored for back-propagation, while keeping the forward activation dense. It is based on the observation that the stored activation that dominates the memory footprint is only employed on back-propagation. Such asymmetric pruning avoids affecting the forward computation, thus making more aggressive pruning possible. Furthermore, we conduct the theoretical analysis for the convergence rate of Back Razor, showing that under mild conditions, our method retains a similar convergence rate as vanilla SGD. Extensive experiments on both Convolutional Neural Networks and Vision Transformers show that Back Razor could yield up to 97% sparsity, saving 9.2x memory usage, without losing accuracy. The code is included in supplementary materials.

## [M<sup>A</sup><sup>3</sup>ViT: Mixture-of-Experts Vision Transformer for Efficient Multi-task Learning with Model-Accelerator Co-design](#)

- hanxue liang · Zhiwen Fan · Rishov Sarkar · Ziyu Jiang · Tianlong Chen · Kai Zou · Yu Cheng · Cong Hao · Zhangyang Wang
- abstract@[open-review](#): Multi-task learning (MTL) encapsulates multiple learned tasks in a single model and often lets those tasks learn better jointly. Multi-tasking models have become successful and often essential for many sophisticated systems such as autonomous driving and indoor robots.

However, when deploying MTL onto those real-world systems that are often resource-constrained or latency-sensitive, two prominent challenges arise: (i) during training, simultaneously optimizing all tasks is often difficult due to gradient conflicts across tasks, and the challenge is amplified when a growing number of tasks have to be squeezed into one compact model; (ii) at inference, current MTL regimes have to activate nearly the entire model even to just execute a single task. Yet most real systems demand only one or two tasks at each moment, while flexibly switching between tasks per need: therefore such “all tasks activated” inference is also highly inefficient and non-scalable in practice. In this paper, we present a model-accelerator co-design framework to enable efficient on-device MTL, that tackles both training and inference bottlenecks. Our framework, dubbed M<sup>3</sup>ViT, customizes mixture-of-experts (MoE) layers into a vision transformer (ViT) backbone for MTL, and sparsely activates task-specific experts during training, which effectively disentangles the parameter spaces to avoid different tasks’ training conflicts. Then at inference with any task of interest, the same design allows for activating only the task-corresponding sparse “expert” pathway, instead of the full model. Our new model design is further enhanced by hardware-level innovations, in particular, a novel computation reordering scheme tailored for memory-constrained MTL that achieves zero-overhead switching between tasks and can scale to any number of experts. Extensive experiments on PASCAL-Context and NYUD-v2 datasets at both software and hardware levels are conducted to demonstrate the effectiveness of the proposed design. When executing the practical scenario of single-task inference, M<sup>3</sup>ViT achieves higher accuracies than encoder-focused MTL methods, while significantly reducing 88% inference FLOPs. When implemented on a hardware platform of one Xilinx ZCU104 FPGA, our co-design framework reduces the memory requirement by 2.40×, while achieving energy efficiency (as the product of latency and power) up to 9.23× times higher than a comparable FPGA baseline.

## [Generalizing Goal-Conditioned Reinforcement Learning with Variational Causal Reasoning](#)

- Wenhao Ding · Haohong Lin · Bo Li · DING ZHAO
- abstract@[open-review](#): As a pivotal component to attaining generalizable solutions in human intelligence, reasoning provides great potential for reinforcement learning (RL) agents' generalization towards varied goals by summarizing part-to-whole arguments and discovering cause-and-effect relations. However, how to discover and represent causalities remains a huge gap that hinders the development of causal RL. In this paper, we augment Goal-Conditioned RL (GCRL) with Causal Graph (CG), a structure built upon the relation between objects and events. We novelly formulate the GCRL problem into variational likelihood maximization with CG as latent variables. To optimize the derived objective, we propose a framework with theoretical performance guarantees that alternates between two steps: using interventional data to estimate the posterior of CG; using CG to learn generalizable models and interpretable policies. Due to the lack of public benchmarks that verify generalization capability under reasoning, we design nine tasks and then empirically show the effectiveness of the proposed method against five baselines on these tasks. Further theoretical analysis shows that our performance improvement is attributed to the virtuous cycle of causal discovery, transition modeling, and policy training, which aligns with the experimental evidence in extensive ablation studies.

## [Near-Optimal Randomized Exploration for Tabular Markov Decision Processes](#)

- Zhihan Xiong · Ruoqi Shen · Qiwen Cui · Maryam Fazel · Simon Du
- abstract@[open-review](#): We study algorithms using randomized value functions for exploration in reinforcement learning. This type of algorithms enjoys appealing empirical performance. We show that when we use 1) a single random seed in each episode, and 2) a Bernstein-type magnitude of noise, we obtain a worst-case  $\widetilde{O}(\sqrt{SAT})$  regret bound for episodic time-inhomogeneous Markov Decision Process where  $S$  is the size of state space,  $A$  is the size of action space,  $H$  is the planning horizon and  $T$  is the number of interactions. This bound polynomially improves all existing bounds for algorithms based on randomized value functions, and for the first time, matches the  $\Omega(\sqrt{SAT})$  lower bound up to logarithmic factors. Our result highlights that randomized exploration can be near-optimal, which was previously achieved only by optimistic algorithms. To achieve the desired result, we develop 1) a new clipping operation to ensure both the probability of being optimistic and the probability of being pessimistic are lower bounded by a constant, and 2) a new recursive formula for the absolute value of estimation errors to analyze the regret.

## [Learning Robust Dynamics through Variational Sparse Gating](#)

- Arnav Kumar Jain · Shivakanth Sujit · Shruti Joshi · Vincent Michalski · Danijar Hafner · Samira Ebrahimi Kahou
- abstract@[open-review](#): Latent dynamics models learn an abstract representation of the environment based on collected experience. For example, world models can imagine unseen trajectories, potentially improving sample efficiency in model-based reinforcement learning. Planning in the real-world requires agents to understand long-term dependencies between actions and events, and account for a varying degree of changes, e.g. due to a change in background or viewpoint. These changes are often quite sparse which suggests incorporating such an inductive bias in a dynamics model. In this work, we introduce Variational Sparse Gating (VSG), where model states are sparsely updated through a stochastic gating mechanism. Moreover, latent state in prior world models comprise of a deterministic and stochastic path and they complement each other for solving tasks. We also propose Simple Variational Sparse Gating (SVSG), which has a purely stochastic latent state and leverages the gating mechanism proposed in VSG. Finally, to evaluate agents in partial-observability and stochasticity, we also introduce a novel environment, called BringBackShapes (BBS). We conducted experiments on BBS and existing benchmarks to demonstrate the benefits of proposed methods.

## [Random Normalization Aggregation for Adversarial Defense](#)

- Minjing Dong · Xinghao Chen · Yunhe Wang · Chang Xu
- abstract@[open-review](#): The vulnerability of deep neural networks has been widely found in various models as well as tasks where slight perturbations on the inputs could lead to incorrect predictions. These perturbed inputs are known as adversarial examples and one of the intriguing properties of them is Adversarial Transfersability, i.e. the capability of adversarial examples to fool other models. Traditionally, this transferability is always regarded as a critical threat to the defense against adversarial attacks, however, we argue that the network robustness can be significantly boosted by utilizing adversarial transferability from a new perspective. In this work, we first discuss the influence of different popular normalization layers on the adversarial transferability, and then provide both empirical evidence and theoretical analysis to shed light on the relationship between normalization types and transferability. Based on our theoretical analysis, we propose a simple yet effective module named Random Normalization Aggregation (RNA) which replaces the batch normalization layers in the networks and aggregates different selected normalization types to form a huge random space. Specifically, a random path is sampled during each inference procedure so that the network itself can be treated as an ensemble of a wide range of different models. Since the entire random space is designed with low adversarial transferability, it is difficult to perform effective attacks even when the network parameters are accessible. We conduct extensive experiments on various models and datasets, and demonstrate the strong superiority of proposed algorithm. The PyTorch code is available at <https://github.com/UniSerj/Random-Norm-Aggregation> and the MindSpore code is available at <https://gitee.com/mindspore/models/tree/master/research/cv/RNA>.

## [BiMLP: Compact Binary Architectures for Vision Multi-Layer Perceptrons](#)

- Yixing Xu · Xinghao Chen · Yunhe Wang
- abstract@[open-review](#): This paper studies the problem of designing compact binary architectures for vision multi-layer perceptrons (MLPs). We provide extensive analysis on the difficulty of binarizing vision MLPs and find that previous binarization methods perform poorly due to limited capacity of binary MLPs. In contrast with the traditional CNNs that utilizing convolutional operations with large kernel size, fully-connected (FC) layers in MLPs can be treated as convolutional layers with kernel size  $1 \times 1$ . Thus, the representation ability of the FC layers will be limited when being binarized, and places restrictions on the capability of spatial mixing and channel mixing on the intermediate features. To this end, we propose to improve the performance of binary MLP (BiMLP) model by enriching the representation ability of binary FC layers. We design a novel binary block that contains multiple branches to merge a series of outputs from the same stage, and also a universal shortcut connection that encourages the information flow from the

previous stage. The downsampling layers are also carefully designed to reduce the computational complexity while maintaining the classification performance. Experimental results on benchmark dataset ImageNet-1k demonstrate the effectiveness of the proposed BiMLP models, which achieve state-of-the-art accuracy compared to prior binary CNNs. The MindSpore code is available at \url{https://gitee.com/mindspore/models/tree/master/research/cv/BiMLP}.

## [Learning Efficient Vision Transformers via Fine-Grained Manifold Distillation](#)

- Zhiwei Hao · Jianyuan Guo · Ding Jia · Kai Han · Yehui Tang · Chao Zhang · Han Hu · Yunhe Wang
- abstract@[open-review](#): In the past few years, transformers have achieved promising performance on various computer vision tasks. Unfortunately, the immense inference overhead of most existing vision transformers withholds them from being deployed on edge devices such as cell phones and smart watches. Knowledge distillation is a widely used paradigm for compressing cumbersome architectures into compact students via transferring information. However, most of them are designed for convolutional neural networks (CNNs), which do not fully investigate the character of vision transformers. In this paper, we fully utilize the patch-level information and propose a fine-grained manifold distillation method for transformer-based networks. Specifically, we train a tiny student model to match a pre-trained teacher model in the patch-level manifold space. Then, we decouple the manifold matching loss into three terms with careful design to further reduce the computational costs for the patch relationship. Equipped with the proposed method, a DeiT-Tiny model containing 5M parameters achieves 76.5% top-1 accuracy on ImageNet-1k, which is +2.0% higher than previous distillation approaches. Transfer learning results on other classification benchmarks and downstream vision tasks also demonstrate the superiority of our method over the state-of-the-art algorithms.

## [Efficient and Modular Implicit Differentiation](#)

- Mathieu Blondel · Quentin Berthet · Marco Cuturi · Roy Frostig · Stephan Hoyer · Felipe Llinares-Lopez · Fabian Pedregosa · Jean-Philippe Vert
- abstract@[open-review](#): Automatic differentiation (autodiff) has revolutionized machine learning. It allows to express complex computations by composing elementary ones in creative ways and removes the burden of computing their derivatives by hand. More recently, differentiation of optimization problem solutions has attracted widespread attention with applications such as optimization layers, and in bi-level problems such as hyper-parameter optimization and meta-learning. However, so far, implicit differentiation remained difficult to use for practitioners, as it often required case-by-case tedious mathematical derivations and implementations. In this paper, we propose automatic implicit differentiation, an efficient and modular approach for implicit differentiation of optimization problems. In our approach, the user defines directly in Python a function `$F$` capturing the optimality conditions of the problem to be differentiated. Once this is done, we leverage autodiff of `$F$` and the implicit function theorem to automatically differentiate the optimization problem. Our approach thus combines the benefits of implicit differentiation and autodiff. It is efficient as it can be added on top of any state-of-the-art solver and modular as the optimality conditions specification is decoupled from the implicit differentiation mechanism. We show that seemingly simple principles allow to recover many existing implicit differentiation methods and create new ones easily. We demonstrate the ease of formulating and solving bi-level optimization problems using our framework. We also showcase an application to the sensitivity analysis of molecular dynamics.

## [Generative Visual Prompt: Unifying Distributional Control of Pre-Trained Generative Models](#)

- Chen Henry Wu · Saman Motamed · Shaunak Srivastava · Fernando D De la Torre
- abstract@[open-review](#): Generative models (e.g., GANs and diffusion models) learn the underlying data distribution in an unsupervised manner. However, many applications of interest require sampling from a specific region of the generative model's output space or even over a range of characteristics. To allow efficient sampling in these scenarios, we propose Generative Visual Prompt (PromptGen), a framework for distributional control over pre-trained generative models by incorporating knowledge of arbitrary off-the-shelf models. PromptGen defines control as an energy-based model (EBM) and samples images in a feed-forward manner by approximating the EBM with invertible neural networks, avoiding optimization at inference. We demonstrate how PromptGen can control several generative models (e.g., StyleGAN2, StyleNeRF, diffusion autoencoder, and NVAE) using various off-the-shelf models: (1) with the CLIP model, PromptGen can sample images guided by text, (2) with image classifiers, PromptGen can de-bias generative models across a set of attributes, and (3) with inverse graphics models, PromptGen can sample images of the same identity in different poses. (4) Finally, PromptGen reveals that the CLIP model shows "reporting bias" when used as control, and PromptGen can further de-bias this controlled distribution in an iterative manner. Our code is available at <https://github.com/ChenWu98/Generative-Visual-Prompt>.

## [Dynamic learning in large matching markets](#)

- Anand Kalvit · Assaf Zeevi
- abstract@[open-review](#): We study a sequential matching problem faced by "large" centralized platforms where "jobs" must be matched to "workers" subject to uncertainty about worker skill proficiencies. Jobs arrive at discrete times with "job-types" observable upon arrival. To capture the "choice overload" phenomenon, we posit an unlimited supply of workers where each worker is characterized by a vector of attributes (aka "worker-types") drawn from an underlying population-level distribution. The distribution as well as mean payoffs for possible worker-job type-pairs are unobservables and the platform's goal is to sequentially match incoming jobs to workers in a way that maximizes its cumulative payoffs over the planning horizon. We establish lower bounds on the "regret" of any matching algorithm in this setting and propose a novel rate-optimal learning algorithm that adapts to aforementioned primitives "online." Our learning guarantees highlight a distinctive characteristic of the problem: achievable performance only has a "second-order" dependence on worker-type distributions; we believe this finding may be of interest more broadly.

## [Movement Penalized Bayesian Optimization with Application to Wind Energy Systems](#)

- Shyam Sundhar Ramesh · Pier Giuseppe Sessa · Andreas Krause · Ilija Bogunovic
- abstract@[open-review](#): Contextual Bayesian optimization (CBO) is a powerful framework for sequential decision-making given side information, with important applications, e.g., in wind energy systems. In this setting, the learner receives context (e.g., weather conditions) at each round, and has to choose an action (e.g., turbine parameters). Standard algorithms assume no cost for switching their decisions at every round. However, in many practical applications, there is a cost associated with such changes, which should be minimized. We introduce the episodic CBO with movement costs problem and, based on the online learning approach for metrical task systems of Coester and Lee (2019), propose a novel randomized mirror descent algorithm that makes use of Gaussian Process confidence bounds. We compare its performance with the offline optimal sequence for each episode and provide rigorous regret guarantees. We further demonstrate our approach on the important real-world application of altitude optimization for Airborne Wind Energy Systems. In the presence of substantial movement costs, our algorithm consistently outperforms standard CBO algorithms.

## [Learning and Covering Sums of Independent Random Variables with Unbounded Support](#)

- Alkis Kalavasis · Konstantinos Stavropoulos · Emmanouil Zampetakis
- abstract@[open-review](#): We study the problem of covering and learning sums  $X = X_1 + \dots + X_n$  of independent integer-valued random variables  $X_i$  (SIIRVs) with infinite support. De et al. at FOCS 2018, showed that even when the collective support of  $X_i$ 's is of size  $4$ , the maximum value of the support necessarily appears in the sample complexity of learning  $X$ . In this work, we address two questions: (i) Are there general families of SIIRVs with infinite support that can be learned with sample complexity independent of both  $n$  and the maximal element of the support? (ii) Are there general families of SIIRVs with infinite support that admit proper sparse covers in total variation distance? As for question (i), we provide a set of simple conditions that allow the infinitely supported SIIRV to be learned with complexity  $\text{poly}(1/\epsilon)$  bypassing the aforementioned lower bound.

We further address question (ii) in the general setting where each variable  $X_i$  has unimodal probability mass function and is a different member of some, possibly multi-parameter, exponential family  $\mathcal{E}$  that satisfies some structural properties. These properties allow  $\mathcal{E}$  to contain heavy tailed and non log-concave distributions. Moreover, we show that for every  $\epsilon > 0$ , and every  $k$ -parameter family  $\mathcal{E}$  that satisfies some structural assumptions, there exists an algorithm with  $\tilde{O}(k \cdot \text{poly}(1/\epsilon))$  samples that learns a sum of  $n$  arbitrary members of  $\mathcal{E}$  within  $\epsilon$  in TV distance. The output of the learning algorithm is also a sum of random variables within the family  $\mathcal{E}$ . En route, we prove that any discrete unimodal exponential family with bounded constant-degree central moments can be approximated by the family corresponding to a bounded subset of the initial (unbounded) parameter space.

## [Manifold Interpolating Optimal-Transport Flows for Trajectory Inference](#)

- Guillaume Huguet · Daniel Sumner Magruder · Oluwadamilola Fasina · Alexander Tong · Manik Kuchroo · Guy Wolf · Smita Krishnaswamy
- abstract@[open-review](#): We present a method called Manifold Interpolating Optimal-Transport Flow (MIOFlow) that learns stochastic, continuous population dynamics from static snapshot samples taken at sporadic timepoints. MIOFlow combines dynamic models, manifold learning, and optimal transport by training neural ordinary differential equations (Neural ODE) to interpolate between static population snapshots as penalized by optimal transport with manifold ground distance. Further, we ensure that the flow follows the geometry by operating in the latent space of an autoencoder that we call a geodesic autoencoder (GAE). In GAE the latent space distance between points is regularized to match a novel multiscale geodesic distance on the data manifold that we define. We show that this method is superior to normalizing flows, Schrödinger bridges and other generative models that are designed to flow from noise to data in terms of interpolating between populations. Theoretically, we link these trajectories with dynamic optimal transport. We evaluate our method on simulated data with bifurcations and merges, as well as scRNA-seq data from embryoid body differentiation, and acute myeloid leukemia treatment.

## [Sample-Efficient Learning of Correlated Equilibria in Extensive-Form Games](#)

- Ziang Song · Song Mei · Yu Bai
- abstract@[open-review](#): Imperfect-Information Extensive-Form Games (IIEFGs) is a prevalent model for real-world games involving imperfect information and sequential plays. The Extensive-Form Correlated Equilibrium (EFCE) has been proposed as a natural solution concept for multi-player general-sum IIEFGs. However, existing algorithms for finding an EFCE require full feedback from the game, and it remains open how to efficiently learn the EFCE in the more challenging bandit feedback setting where the game can only be learned by observations from repeated playing. This paper presents the first sample-efficient algorithm for learning the EFCE from bandit feedback. We begin by proposing  $K$ -EFCE---a generalized definition that allows players to observe and deviate from the recommended actions for  $K$  times. The  $K$ -EFCE includes the EFCE as a special case at  $K=1$ , and is an increasingly stricter notion of equilibrium as  $K$  increases. We then design an uncoupled no-regret algorithm that finds an  $\epsilon$ -approximate  $K$ -EFCE within  $\tilde{O}(\max_i X_i A_i^K / \epsilon^2)$  iterations in the full feedback setting, where  $X_i$  and  $A_i$  are the number of information sets and actions for the  $i$ -th player. Our algorithm works by minimizing a wide-range regret at each information set that takes into account all possible recommendation histories. Finally, we design a sample-based variant of our algorithm that learns an  $\epsilon$ -approximate  $K$ -EFCE within  $\tilde{O}(\max_i X_i A_i^{K+1} / \epsilon^2)$  episodes of play in the bandit feedback setting. When specialized to  $K=1$ , this gives the first sample-efficient algorithm for learning EFCE from bandit feedback.

## [Fast Bayesian Coresets via Subsampling and Quasi-Newton Refinement](#)

- Cian Naik · Judith Rousseau · Trevor Campbell
- abstract@[open-review](#): Bayesian coresets approximate a posterior distribution by building a small weighted subset of the data points. Any inference procedure that is too computationally expensive to be run on the full posterior can instead be run inexpensively on the coreset, with results that approximate those on the full data. However, current approaches are limited by either a significant run-time or the need for the user to specify a low-cost approximation to the full posterior. We propose a Bayesian coreset construction algorithm that first selects a uniformly random subset of data, and then optimizes the weights using a novel quasi-Newton method. Our algorithm is a simple to implement, black-box method, that does not require the user to specify a low-cost posterior approximation. It is the first to come with a general high-probability bound on the KL divergence of the output coreset posterior. Experiments demonstrate that our method provides significant improvements in coreset quality against alternatives with comparable construction times, with far less storage cost and user input required.

## [Instance-Based Uncertainty Estimation for Gradient-Boosted Regression Trees](#)

- Jonathan Brophy · Daniel Lowd
- abstract@[open-review](#): Gradient-boosted regression trees (GBRTs) are hugely popular for solving tabular regression problems, but provide no estimate of uncertainty. We propose Instance-Based Uncertainty estimation for Gradient-boosted regression trees (IBUG), a simple method for extending any GBRT point predictor to produce probabilistic predictions. IBUG computes a non-parametric distribution around a prediction using the k-nearest training instances, where distance is measured with a tree-ensemble kernel. The runtime of IBUG depends on the number of training examples at each leaf in the ensemble, and can be improved by sampling trees or training instances. We also find that IBUG can achieve improved probabilistic performance by using different base GBRT models, and can more flexibly model the posterior distribution of a prediction than competing methods. We also find that previous methods suffer from poor probabilistic calibration on some datasets, which can be mitigated using a scalar factor tuned on the validation data.

## [Scalable Representation Learning in Linear Contextual Bandits with Constant Regret Guarantees](#)

- Andrea Tirinzoni · Matteo Papini · Ahmed Touati · Alessandro Lazaric · Matteo Pirotta
- abstract@[open-review](#): We study the problem of representation learning in stochastic contextual linear bandits. While the primary concern in this domain is usually to find realizable representations (i.e., those that allow predicting the reward function at any context-action pair exactly), it has been recently shown that representations with certain spectral properties (called HLS) may be more effective for the exploration-exploitation task, enabling LinUCB to achieve constant (i.e., horizon-independent) regret. In this paper, we propose BanditSRL, a representation learning algorithm that combines a novel constrained optimization problem to learn a realizable representation with good spectral properties with a generalized likelihood ratio test to exploit the recovered representation and avoid excessive exploration. We prove that BanditSRL can be paired with any no-regret algorithm and achieve constant regret whenever an HLS representation is available. Furthermore, BanditSRL can be easily combined with deep neural networks and we show how regularizing towards HLS representations is beneficial in standard benchmarks.

## [Near Instance-Optimal PAC Reinforcement Learning for Deterministic MDPs](#)

- Andrea Tirinzoni · Aymen Al Marjani · Emilie Kaufmann
- abstract@[open-review](#): In probably approximately correct (PAC) reinforcement learning (RL), an agent is required to identify an  $\epsilon$ -optimal policy with probability  $1 - \delta$ . While minimax optimal algorithms exist for this problem, its instance-dependent complexity remains elusive in episodic Markov decision processes (MDPs). In this paper, we propose the first (nearly) matching upper and lower bounds on the sample complexity of PAC RL in deterministic episodic MDPs with finite state and action spaces. In particular, our bounds feature a new notion of sub-optimality gap for state-action pairs that we call the deterministic return gap. While our instance-dependent lower bound is written as a linear program, our algorithms are very simple and do not require solving such an optimization problem during learning. Their design and analyses employ novel ideas, including graph-theoretical concepts (minimum flows) and a new maximum-coverage exploration strategy.

## [Global Linear and Local Superlinear Convergence of IRLS for Non-Smooth Robust Regression](#)

- Liangzu Peng · Christian Kämmerle · René Vidal
- abstract@[open-review](#): We advance both the theory and practice of robust  $\ell_p$ -quasinorm regression for  $p \in (0,1]$  by using novel variants of iteratively reweighted least-squares (IRLS) to solve the underlying non-smooth problem. In the convex case,  $p=1$ , we prove that this IRLS variant converges globally at a linear rate under a mild, deterministic condition on the feature matrix called the stable range space property. In the non-convex case,  $p \in (0,1)$ , we prove that under a similar condition, IRLS converges locally to the global minimizer at a superlinear rate of order  $2-p$ ; the rate becomes quadratic as  $p \rightarrow 0$ . We showcase the proposed methods in three applications: real phase retrieval, regression without correspondences, and robust face restoration. The results show that (1) IRLS can handle a larger number of outliers than other methods, (2) it is faster than competing methods at the same level of accuracy, (3) it restores a sparsely corrupted face image with satisfactory visual quality.

## [Improved Differential Privacy for SGD via Optimal Private Linear Operators on Adaptive Streams](#)

- Sergey Denisov · H. Brendan McMahan · John Rush · Adam Smith · Abhradeep Guha Thakurta
- abstract@[open-review](#): Motivated by recent applications requiring differential privacy in the setting of adaptive streams, we investigate the question of optimal instantiations of the matrix mechanism in this setting. We prove fundamental theoretical results on the applicability of matrix factorizations to the adaptive streaming setting, and provide a new parameter-free fixed-point algorithm for computing optimal factorizations. We instantiate this framework with respect to concrete matrices which arise naturally in the machine learning setting, and train user-level differentially private models with the resulting optimal mechanisms, yielding significant improvements on a notable problem in federated learning with user-level differential privacy.

## [Finite-Time Regret of Thompson Sampling Algorithms for Exponential Family Multi-Armed Bandits](#)

- Tianyuan Jin · Pan Xu · Xiaokui Xiao · Anima Anandkumar
- abstract@[open-review](#): We study the regret of Thompson sampling (TS) algorithms for exponential family bandits, where the reward distribution is from a one-dimensional exponential family, which covers many common reward distributions including Bernoulli, Gaussian, Gamma, Exponential, etc. We propose a Thompson sampling algorithm, termed ExpTS, which uses a novel sampling distribution to avoid the under-estimation of the optimal arm. We provide a tight regret analysis for ExpTS, which simultaneously yields both the finite-time regret bound as well as the asymptotic regret bound. In particular, for a  $K$ -armed bandit with exponential family rewards, ExpTS over a horizon  $T$  is sub-UCB (a strong criterion for the finite-time regret that is problem-dependent), minimax optimal up to a factor  $\sqrt{\log K}$ , and asymptotically optimal, for exponential family rewards. Moreover, we propose ExpTS<sup>++</sup>, by adding a greedy exploitation step in addition to the sampling distribution used in ExpTS, to avoid the over-estimation of sub-optimal arms. ExpTS<sup>++</sup> is an anytime bandit algorithm and achieves the minimax optimality and asymptotic optimality simultaneously for exponential family reward distributions. Our proof techniques are general and conceptually simple and can be easily applied to analyze standard Thompson sampling with specific reward distributions.

## [Independence Testing-Based Approach to Causal Discovery under Measurement Error and Linear Non-Gaussian Models](#)

- Haoyue Dai · Peter Spirtes · Kun Zhang
- abstract@[open-review](#): Causal discovery aims to recover causal structures generating the observational data. Despite its success in certain problems, in many real-world scenarios the observed variables are not the target variables of interest, but the imperfect measures of the target variables. Causal discovery under measurement error aims to recover the causal graph among unobserved target variables from observations made with measurement error. We consider a specific formulation of the problem, where the unobserved target variables follow a linear non-Gaussian acyclic model, and the measurement process follows the random measurement error model. Existing methods on this formulation rely on non-scalable over-complete independent component analysis (OICA). In this work, we propose the Transformed Independent Noise (TIN) condition, which checks for independence between a specific linear transformation of some measured variables and certain other measured variables. By leveraging the non-Gaussianity and higher-order statistics of data, TIN is informative about the graph structure among the unobserved target variables. By utilizing TIN, the ordered group decomposition of the causal model is identifiable. In other words, we could achieve what once required OICA to achieve by only conducting independence tests. Experimental results on both synthetic and real-world data demonstrate the effectiveness and reliability of our method.

## [Biologically-Plausible Determinant Maximization Neural Networks for Blind Separation of Correlated Sources](#)

- Bariscan Bozkurt · Cengiz Pehlevan · Alper Erdogan
- abstract@[open-review](#): Extraction of latent sources of complex stimuli is critical for making sense of the world. While the brain solves this blind source separation (BSS) problem continuously, its algorithms remain unknown. Previous work on biologically-plausible BSS algorithms assumed that observed signals are linear mixtures of statistically independent or uncorrelated sources, limiting the domain of applicability of these algorithms. To overcome this limitation, we propose novel biologically-plausible neural networks for the blind separation of potentially dependent/correlated sources. Differing from previous work, we assume some general geometric, not statistical, conditions on the source vectors allowing separation of potentially dependent/correlated sources. Concretely, we assume that the source vectors are sufficiently scattered in their domains which can be described by certain polytopes. Then, we consider recovery of these sources by the Det-Max criterion, which maximizes the determinant of the output correlation matrix to enforce a similar spread for the source estimates. Starting from this normative principle, and using a weighted similarity matching approach that enables arbitrary linear transformations adaptable by local learning rules, we derive two-layer biologically-plausible neural network algorithms that can separate mixtures into sources coming from a variety of source domains. We demonstrate that our algorithms outperform other biologically-plausible BSS algorithms on correlated source separation problems.

## [Graph Few-shot Learning with Task-specific Structures](#)

- Song Wang · Chen Chen · Jundong Li
- abstract@[open-review](#): Graph few-shot learning is of great importance among various graph learning tasks. Under the few-shot scenario, models are required to conduct classification given limited labeled samples. Existing graph few-shot learning methods typically leverage Graph Neural Networks (GNNs) and perform classification across a series of meta-tasks. Nevertheless, these methods generally rely on the original graph (i.e., the graph that the meta-task is sampled from) to learn node representations. Consequently, the learned representations for the same nodes are identical in all meta-tasks. Since the class sets are different across meta-tasks, node representations should be task-specific to promote classification performance. Therefore, to adaptively learn node representations across meta-tasks, we propose a novel framework that learns a task-specific structure for each meta-task. To handle the variety of nodes across meta-tasks, we extract relevant nodes and learn task-specific structures based on node influence and mutual information. In this way, we can learn node representations with the task-specific structure tailored for each meta-task. We further conduct extensive experiments on five node classification datasets under both single- and multiple-graph settings to validate the superiority of our framework over the state-of-the-art baselines.

## [Self-Consistent Dynamical Field Theory of Kernel Evolution in Wide Neural Networks](#)

- Blake Bordelon · Cengiz Pehlevan
- abstract@[open-review](#): We analyze feature learning in infinite-width neural networks trained with gradient flow through a self-consistent dynamical field theory. We construct a collection of deterministic dynamical order parameters which are inner-product kernels for hidden unit activations and gradients in

each layer at pairs of time points, providing a reduced description of network activity through training. These kernel order parameters collectively define the hidden layer activation distribution, the evolution of the neural tangent kernel, and consequently output predictions. We show that the field theory derivation recovers the recursive stochastic process of infinite-width feature learning networks obtained from Yang & Hu with Tensor Programs. For deep linear networks, these kernels satisfy a set of algebraic matrix equations. For nonlinear networks, we provide an alternating sampling procedure to self-consistently solve for the kernel order parameters. We provide comparisons of the self-consistent solution to various approximation schemes including the static NTK approximation, gradient independence assumption, and leading order perturbation theory, showing that each of these approximations can break down in regimes where general self-consistent solutions still provide an accurate description. Lastly, we provide experiments in more realistic settings which demonstrate that the loss and kernel dynamics of CNNs at fixed feature learning strength is preserved across different widths on a CIFAR classification task.

## [Extrapolative Continuous-time Bayesian Neural Network for Fast Training-free Test-time Adaptation](#)

- Henguan Huang · Xiangming Gu · Hao Wang · Chang Xiao · Hongfu Liu · Ye Wang
- abstract@[open-review](#): Human intelligence has shown remarkably lower latency and higher precision than most AI systems when processing non-stationary streaming data in real-time. Numerous neuroscience studies suggest that such abilities may be driven by internal predictive modeling. In this paper, we explore the possibility of introducing such a mechanism in unsupervised domain adaptation (UDA) for handling non-stationary streaming data for real-time streaming applications. We propose to formulate internal predictive modeling as a continuous-time Bayesian filtering problem within the context of a stochastic dynamical system. Such a dynamical system describes the dynamics of model parameters of a UDA model evolving with non-stationary streaming data. Building on such a dynamical system, we then develop extrapolative continuous-time Bayesian neural networks (ECBNN), which generalize existing Bayesian neural networks to represent temporal dynamics and allow us to extrapolate the distribution of model parameters before observing the incoming data, therefore effectively reducing the latency. Remarkably, our empirical results show that ECBNN is capable of continuously generating better distributions of model parameters along the time axis given historical data only, thereby achieving (1) training-free online adaptation with low latency, (2) gradually improved alignment between the source and target features and (3) gradually improved model performance over time during the real-time testing stage.

## [Learning Generalizable Part-based Feature Representation for 3D Point Clouds](#)

- Xin Wei · Xiang Gu · Jian Sun
- abstract@[open-review](#): Deep networks on 3D point clouds have achieved remarkable success in 3D classification, while they are vulnerable to geometry variations caused by inconsistent data acquisition procedures. This results in a challenging 3D domain generalization (3DDG) problem, that is to generalize a model trained on source domain to an unseen target domain. Based on the observation that local geometric structures are more generalizable than the whole shape, we propose to reduce the geometry shift by a generalizable part-based feature representation and design a novel part-based domain generalization network (PDG) for 3D point cloud classification. Specifically, we build a part-template feature space shared by source and target domains. Shapes from distinct domains are first organized to part-level features and then represented by part-template features. The transformed part-level features, dubbed aligned part-based representations, are then aggregated by a part-based feature aggregation module. To improve the robustness of the part-based representations, we further propose a contrastive learning framework upon part-based shape representation. Experiments and ablation studies on 3DDG benchmarks justify the efficacy of the proposed approach for domain generalization, compared with the previous state-of-the-art methods.

## [Invariance Learning based on Label Hierarchy](#)

- Shoji Toyota · Kenji Fukumizu
- abstract@[open-review](#): Deep Neural Networks inherit spurious correlations embedded in training data and hence may fail to predict desired labels on unseen domains (or environments), which have different distributions from the domain to provide training data. Invariance Learning (IL) has been developed recently to overcome this shortcoming; using training data in many domains, IL estimates such a predictor that is invariant to a change of domain. However, the requirement of training data in multiple domains is a strong restriction of using IL, since it demands expensive annotation. We propose a novel IL framework to overcome this problem. Assuming the availability of data from multiple domains for a higher level of classification task, for which the labeling cost is lower, we estimate an invariant predictor for the target classification task with training data gathered in a single domain. Additionally, we propose two cross-validation methods for selecting hyperparameters of invariance regularization, which has not been addressed properly in existing IL methods. The effectiveness of the proposed framework, including the cross-validation, is demonstrated empirically. Theoretical analysis reveals that our framework can estimate the desirable invariant predictor with a hyperparameter fixed correctly, and that such a preferable hyperparameter is chosen by the proposed CV methods under some conditions.

## [TVLT: Textless Vision-Language Transformer](#)

- Zineng Tang · Jaemin Cho · Yixin Nie · Mohit Bansal
- abstract@[open-review](#): Long before the emergence of written symbols, speech has been the main modality of verbal communication among humans. In this work, we present the Textless Vision-Language Transformer (TVLT), a transformer model that takes raw audio and visual inputs for vision-and-language representation learning with minimal modality-specific design, and does not use extra text-specific modules such as tokenization or automatic speech recognition (ASR). We show that TVLT attains results comparable to its text-based counterpart, on various multi-modal tasks such as video retrieval, image retrieval, visual question answering, and multimodal sentiment analysis, while being 50x faster during inference and with only one-third of the parameters. Our findings suggest the promising possibility of obtaining compact and efficient visual-linguistic representations by learning directly from low-level visual and audio perception signals.

## [GENIE: Higher-Order Denoising Diffusion Solvers](#)

- Tim Dockhorn · Arash Vahdat · Karsten Kreis
- abstract@[open-review](#): Denoising diffusion models (DDMs) have emerged as a powerful class of generative models. A forward diffusion process slowly perturbs the data, while a deep model learns to gradually denoise. Synthesis amounts to solving a differential equation (DE) defined by the learnt model. Solving the DE requires slow iterative solvers for high-quality generation. In this work, we propose Higher-Order Denoising Diffusion Solvers (GENIE): Based on truncated Taylor methods, we derive a novel higher-order solver that significantly accelerates synthesis. Our solver relies on higher-order gradients of the perturbed data distribution, that is, higher-order score functions. In practice, only Jacobian-vector products (JVPs) are required and we propose to extract them from the first-order score network via automatic differentiation. We then distill the JVPs into a separate neural network that allows us to efficiently compute the necessary higher-order terms for our novel sampler during synthesis. We only need to train a small additional head on top of the first-order score network. We validate GENIE on multiple image generation benchmarks and demonstrate that GENIE outperforms all existing solvers. Unlike recent methods that fundamentally alter the generation process in DDMs, our GENIE solves the true generative DE and still enables applications such as encoding and guided sampling.

## [Neural Attentive Circuits](#)

- Martin Weiss · Nasim Rahaman · Francesco Locatello · Chris Pal · Yoshua Bengio · Bernhard Schölkopf · Li Erran Li · Nicolas Ballas
- abstract@[open-review](#): Recent work has seen the development of general purpose neural architectures that can be trained to perform tasks across diverse data modalities. General purpose models typically make few assumptions about the underlying data-structure and are known to perform well in the large-

data regime. At the same time, there has been growing interest in modular neural architectures that represent the data using sparsely interacting modules. These models can be more robust out-of-distribution, computationally efficient, and capable of sample-efficient adaptation to new data. However, they tend to make domain-specific assumptions about the data, and present challenges in how module behavior (i.e., parameterization) and connectivity (i.e., their layout) can be jointly learned. In this work, we introduce a general purpose, yet modular neural architecture called Neural Attentive Circuits (NACs) that jointly learns the parameterization and a sparse connectivity of neural modules without using domain knowledge. NACs are best understood as the combination of two systems that are jointly trained end-to-end: one that determines the module configuration and the other that executes it on an input. We demonstrate qualitatively that NACs learn diverse and meaningful module configurations on the Natural Language and Visual Reasoning for Real (NLVR2) dataset without additional supervision. Quantitatively, we show that by incorporating modularity in this way, NACs improve upon a strong non-modular baseline in terms of low-shot adaptation on CIFAR and Caltech-UCSD Birds dataset (CUB) by  $\sim 10\%$ , and OOD robustness on Tiny ImageNet-R by  $\sim 2.5\%$ , and we also find that NACs can achieve an 8x speedup at inference time while losing less than 3% performance.

## [Improving Intrinsic Exploration with Language Abstractions](#)

- Jesse Mu · Victor Zhong · Roberta Raileanu · Minqi Jiang · Noah Goodman · Tim Rocktäschel · Edward Grefenstette
- abstract@[open-review](#): Reinforcement learning (RL) agents are particularly hard to train when rewards are sparse. One common solution is to use intrinsic rewards to encourage agents to explore their environment. However, recent intrinsic exploration methods often use state-based novelty measures which reward low-level exploration and may not scale to domains requiring more abstract skills. Instead, we explore natural language as a general medium for highlighting relevant abstractions in an environment. Unlike previous work, we evaluate whether language can improve over existing exploration methods by directly extending (and comparing to) competitive intrinsic exploration baselines: AMIGo (Campero et al., 2021) and NovelID (Zhang et al., 2021). These language-based variants outperform their non-linguistic forms by 23-46% across 13 challenging tasks from the MiniGrid and MiniHack environment suites.

## [A Projection-free Algorithm for Constrained Stochastic Multi-level Composition Optimization](#)

- Tesi Xiao · Krishnakumar Balasubramanian · Saeed Ghadimi
- abstract@[open-review](#): We propose a projection-free conditional gradient-type algorithm for smooth stochastic multi-level composition optimization, where the objective function is a nested composition of  $T$  functions and the constraint set is a closed convex set. Our algorithm assumes access to noisy evaluations of the functions and their gradients, through a stochastic first-order oracle satisfying certain standard unbiasedness and second-moment assumptions. We show that the number of calls to the stochastic first-order oracle and the linear-minimization oracle required by the proposed algorithm, to obtain an  $\epsilon$ -stationary solution, are of order  $\mathcal{O}_T(\epsilon^{-2})$  and  $\mathcal{O}_T(\epsilon^{-3})$  respectively, where  $\mathcal{O}_T$  hides constants in  $T$ . Notably, the dependence of these complexity bounds on  $\epsilon$  and  $T$  are separate in the sense that changing one does not impact the dependence of the bounds on the other. For the case of  $T=1$ , we also provide a high-probability convergence result that depends poly-logarithmically on the inverse confidence level. Moreover, our algorithm is parameter-free and does not require any (increasing) order of mini-batches to converge unlike the common practice in the analysis of stochastic conditional gradient-type algorithms.

## [Theseus: A Library for Differentiable Nonlinear Optimization](#)

- Luis Pineda · Taosha Fan · Maurizio Monge · Shobha Venkataraman · Paloma Sodhi · Ricky T. Q. Chen · Joseph Ortiz · Daniel DeTone · Austin Wang · Stuart Anderson · Jing Dong · Brandon Amos · Mustafa Mukadam
- abstract@[open-review](#): We present Theseus, an efficient application-agnostic open source library for differentiable nonlinear least squares (DNLS) optimization built on PyTorch, providing a common framework for end-to-end structured learning in robotics and vision. Existing DNLS implementations are application specific and do not always incorporate many ingredients important for efficiency. Theseus is application-agnostic, as we illustrate with several example applications that are built using the same underlying differentiable components, such as second-order optimizers, standard costs functions, and Lie groups. For efficiency, Theseus incorporates support for sparse solvers, automatic vectorization, batching, GPU acceleration, and gradient computation with implicit differentiation and direct loss minimization. We do extensive performance evaluation in a set of applications, demonstrating significant efficiency gains and better scalability when these features are incorporated.

## [OPEN: Orthogonal Propagation with Ego-Network Modeling](#)

- Liang Yang · Lina Kang · Qiuliang Zhang · Mengzhe Li · bingxin niu · Dongxiao He · Zhen Wang · Chuan Wang · Xiaochun Cao · Yuanfang Guo
- abstract@[open-review](#): To alleviate the unfavorable effect of noisy topology in Graph Neural networks (GNNs), some efforts perform the local topology refinement through the pairwise propagation weight learning and the multi-channel extension. Unfortunately, most of them suffer a common and fatal drawback: irrelevant propagation to one node and in multi-channels. These two kinds of irrelevances make propagation weights in multi-channels free to be determined by the labeled data, and thus the GNNs are exposed to overfitting. To tackle this issue, a novel Orthogonal Propagation with Ego-Network modeling (OPEN) is proposed by modeling relevances between propagations. Specifically, the relevance between propagations to one node is modeled by whole ego-network modeling, while the relevance between propagations in multi-channels is modeled via diversity requirement. By interpreting the propagations to one node from the perspective of dimension reduction, propagation weights are inferred from principal components of the ego-network, which are orthogonal to each other. Theoretical analysis and experimental evaluations reveal four attractive characteristics of OPEN as modeling high-order relationships beyond pairwise one, preventing overfitting, robustness, and high efficiency.

## [On the Learning Mechanisms in Physical Reasoning](#)

- Shiqian Li · Kewen Wu · Chi Zhang · Yixin Zhu
- abstract@[open-review](#): Is dynamics prediction indispensable for physical reasoning? If so, what kind of roles do the dynamics prediction modules play during the physical reasoning process? Most studies focus on designing dynamics prediction networks and treating physical reasoning as a downstream task without investigating the questions above, taking for granted that the designed dynamics prediction would undoubtedly help the reasoning process. In this work, we take a closer look at this assumption, exploring this fundamental hypothesis by comparing two learning mechanisms: Learning from Dynamics (LfD) and Learning from Intuition (LfI). In the first experiment, we directly examine and compare these two mechanisms. Results show a surprising finding: Simple LfI is better than or on par with state-of-the-art LfD. This observation leads to the second experiment with Ground-truth Dynamics (GD), the ideal case of LfD wherein dynamics are obtained directly from a simulator. Results show that dynamics, if directly given instead approximated, would achieve much higher performance than LfI alone on physical reasoning; this essentially serves as the performance upper bound. Yet practically, LfD mechanism can only predict Approximate Dynamics (AD) using dynamics learning modules that mimic the physical laws, making the following downstream physical reasoning modules degenerate into the LfI paradigm; see the third experiment. We note that this issue is hard to mitigate, as dynamics prediction errors inevitably accumulate in the long horizon. Finally, in the fourth experiment, we note that LfI, the extremely simpler strategy when done right, is more effective in learning to solve physical reasoning problems. Taken together, the results on the challenging benchmark of PHYRE [3] show that LfI is, if not better, as good as LfD with bells and whistles for dynamics prediction. However, the potential improvement from LfD, though challenging, remains lucrative.

## [BinauralGrad: A Two-Stage Conditional Diffusion Probabilistic Model for Binaural Audio Synthesis](#)

- Yichong Leng · Zehua Chen · Junliang Guo · Haohe Liu · Jiawei Chen · Xu Tan · Danilo Mandic · Lei He · Xiangyang Li · Tao Qin · Sheng Zhao · Tie-Yan Liu
- abstract@[open-review](#): Binaural audio plays a significant role in constructing immersive augmented and virtual realities. As it is expensive to record binaural audio from the real world, synthesizing them from mono audio has attracted increasing attention. This synthesis process involves not only the basic physical warping of the mono audio, but also room reverberations and head/ear related filtration, which, however, are difficult to accurately simulate in traditional digital signal processing. In this paper, we formulate the synthesis process from a different perspective by decomposing the binaural audio into a common part that shared by the left and right channels as well as a specific part that differs in each channel. Accordingly, we propose BinauralGrad, a novel two-stage framework equipped with diffusion models to synthesize them respectively. Specifically, in the first stage, the common information of the binaural audio is generated with a single-channel diffusion model conditioned on the mono audio, based on which the binaural audio is generated by a two-channel diffusion model in the second stage. Combining this novel perspective of two-stage synthesis with advanced generative models (i.e., the diffusion models), the proposed BinauralGrad is able to generate accurate and high-fidelity binaural audio samples. Experiment results show that on a benchmark dataset, BinauralGrad outperforms the existing baselines by a large margin in terms of both object and subject evaluation metrics (Wave L2: \$0.128\$ vs. \$0.157\$, MOS: \$3.80\$ vs. \$3.61\$). The generated audio samples are available online\footnote{\url{https://speechresearch.github.io/binauralgrad}}.

## [VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training](#)

- Zhan Tong · Yibing Song · Jue Wang · Limin Wang
- abstract@[open-review](#): Pre-training video transformers on extra large-scale datasets is generally required to achieve premier performance on relatively small datasets. In this paper, we show that video masked autoencoders (VideoMAE) are data-efficient learners for self-supervised video pre-training (SSVP). We are inspired by the recent ImageMAE and propose customized video tube masking with an extremely high ratio. This simple design makes video reconstruction a more challenging and meaningful self-supervision task, thus encouraging extracting more effective video representations during this pre-training process. We obtain three important findings on SSVP: (1) An extremely high proportion of masking ratio (i.e., 90% to 95%) still yields favorable performance of VideoMAE. The temporally redundant video content enables higher masking ratio than that of images. (2) VideoMAE achieves impressive results on very small datasets (i.e., around 3k-4k videos) without using any extra data. This is partially ascribed to the challenging task of video reconstruction to enforce high-level structure learning. (3) VideoMAE shows that data quality is more important than data quantity for SSVP. Domain shift between pre-training and target datasets is an important issue. Notably, our VideoMAE with the vanilla ViT backbone can achieve 84.7% on Kinetics-400, 75.3% on Something-Something V2, 90.8% on UCF101, and 61.1% on HMDB51, without using any extra data.

## [Explaining Graph Neural Networks with Structure-Aware Cooperative Games](#)

- Shichang Zhang · Neil Shah · Yozen Liu · Yizhou Sun
- abstract@[open-review](#): Explaining predictions made by machine learning models is important and has attracted increased interest. The Shapley value from cooperative game theory has been proposed as a prime approach for computing feature importance towards predictions, especially for images, text, tabular data, and recently graph neural networks (GNNs) on graphs. In this work, we revisit the appropriateness of the Shapley value for graph explanation, where the task is to identify the most important subgraph and constituent nodes for graph-level predictions. We purport that the Shapley value is a non-ideal choice for graph data because it is by definition not structure-aware. We propose a Graph Structure-aware eXplanation (GStarX) method to leverage the critical graph structure information to improve the explanation. Specifically, we propose a scoring function based on a new structure-aware value from the cooperative game theory called the HN value. When used to score node importance, the HN value utilizes graph structures to attribute cooperation surplus between neighbor nodes, resembling message passing in GNNs, so that node importance scores reflect not only the node feature importance but also the structural roles. We demonstrate that GStarX produces qualitatively more intuitive explanations, and quantitatively improves over strong baselines on chemical graph property prediction, text graph sentiment classification, and synthetic subgraph detection tasks.

## [EcoFormer: Energy-Saving Attention with Linear Complexity](#)

- Jing Liu · Zizheng Pan · Haoyu He · Jianfei Cai · Bohan Zhuang
- abstract@[open-review](#): Transformer is a transformative framework for deep learning which models sequential data and has achieved remarkable performance on a wide range of tasks, but with high computational and energy cost. To improve its efficiency, a popular choice is to compress the models via binarization which constrains the floating-point values into binary ones to significantly save resource consumption owing to cheap bitwise operations. However, existing binarization methods only aim at minimizing the information loss for the input distribution statistically, while ignoring the pairwise similarity modeling at the core of the attention mechanism. To this end, we propose a new binarization paradigm customized to high-dimensional softmax attention via kernelized hashing, called EcoFormer, to map the original queries and keys into low-dimensional binary codes in Hamming space. The kernelized hash functions are learned to match the ground-truth similarity relations extracted from the attention map in a self-supervised way. Based on the equivalence between the inner product of binary codes and the Hamming distance as well as the associative property of matrix multiplication, we can approximate the attention in linear complexity by expressing it as a dot-product of binary codes. Moreover, the compact binary representations of queries and keys in EcoFormer enable us to replace most of the expensive multiply-accumulate operations in attention with simple accumulations to save considerable on-chip energy footprint on edge devices. Extensive experiments on both vision and language tasks show that EcoFormer consistently achieves comparable performance with standard attentions while consuming much less resources. For example, based on PVTv2-B0 and ImageNet-1K, EcoFormer achieves 73% reduction in energy footprint with only a slight performance drop of 0.33% compared to the standard attention.

## [Deep Multi-Modal Structural Equations For Causal Effect Estimation With Unstructured Proxies](#)

- Shachi Deshpande · Kaiwen Wang · Dhruv Sreenivas · Zheng Li · Volodymyr Kuleshov
- abstract@[open-review](#): Estimating the effect of an intervention while accounting for confounding variables is a key task in causal inference. Oftentimes, the confounders are unobserved, but we have access to large amounts of unstructured data (images, text) that contain valuable proxy signal about the missing confounders. This paper demonstrates that leveraging unstructured data that is typically left unused by existing algorithms improves the accuracy of causal effect estimation. Specifically, we introduce deep multi-modal structural equations, a generative model in which confounders are latent variables and unstructured data are proxy variables. This model supports multiple multi-modal proxies (images, text) as well as missing data. We empirically demonstrate on tasks in genomics and healthcare that our approach corrects for confounding using unstructured inputs, potentially enabling the use of large amounts of data that were previously not used in causal inference.

## [Efficient coding, channel capacity, and the emergence of retinal mosaics](#)

- Na Young Jun · Greg Field · John Pearson
- abstract@[open-review](#): Among the most striking features of retinal organization is the grouping of its output neurons, the retinal ganglion cells, into a diversity of functional types. Each of these types exhibits a mosaic-like organization of receptive fields that tiles the retina and visual space. Previous work has shown that many features of retinal ganglion cell organization, including the existence of ON and OFF cell types, the structure of spatial receptive fields, and their relative arrangement, can be predicted on the basis of efficient coding theory. This theory posits that the nervous system is organized to maximize information in its encoding of stimuli while minimizing metabolic costs. Here, we use efficient coding theory to present a comprehensive account of mosaic organization in the case of natural images as the retinal channel capacity—the number of neurons available for encoding—is varied. Using a simple model of efficient coding, we show that, beginning with spatially localized temporal smoothing filters, mosaic density increases with channel capacity up to a series of critical points at which new mosaics focused on increasingly high temporal frequencies emerge. In addition, we show

both experimentally and theoretically that a transition from alignment to anti-alignment is observed not only with increasing output noise as previously reported, but also with decreasing input noise. Together, these results offer a unified perspective on the relationship between retinal mosaics, efficient coding, and channel capacity that may help to explain the stunning functional diversity of retinal mosaics.

## [Learning \(Very\) Simple Generative Models Is Hard](#)

- Sitan Chen · Jerry Li · Yuanzhi Li
- abstract@[open-review](#): Motivated by the recent empirical successes of deep generative models, we study the computational complexity of the following unsupervised learning problem. For an unknown neural network  $F: \mathbb{R}^d \rightarrow \mathbb{R}^{d'}$ , let  $D$  be the distribution over  $\mathbb{R}^{d'}$  given by pushing the standard Gaussian  $\mathcal{N}(0, \text{Id}_d)$  through  $F$ . Given i.i.d. samples from  $D$ , the goal is to output *any* distribution close to  $D$  in statistical distance. We show under the statistical query (SQ) model that no polynomial-time algorithm can solve this problem even when the output coordinates of  $F$  are one-hidden-layer ReLU networks with  $\log(d)$  neurons. Previously, the best lower bounds for this problem simply followed from lower bounds for *supervised learning* and required at least two hidden layers and  $\text{poly}(d)$  neurons [Daniely-Vardi '21, Chen-Gollakota-Klivans-Meka '22]. The key ingredient in our proof is an ODE-based construction of a compactly supported, piecewise-linear function  $f$  with polynomially-bounded slopes such that the pushforward of  $\mathcal{N}(0, I_d)$  under  $f$  matches all low-degree moments of  $\mathcal{N}(0, I_d)$ .

## [Self-Supervised Visual Representation Learning with Semantic Grouping](#)

- Xin Wen · Bingchen Zhao · Anlin Zheng · Xiangyu Zhang · Xiaojuan Qi
- abstract@[open-review](#): In this paper, we tackle the problem of learning visual representations from unlabeled scene-centric data. Existing works have demonstrated the potential of utilizing the underlying complex structure within scene-centric data; still, they commonly rely on hand-crafted objectness priors or specialized pretext tasks to build a learning framework, which may harm generalizability. Instead, we propose contrastive learning from data-driven semantic slots, namely SlotCon, for joint semantic grouping and representation learning. The semantic grouping is performed by assigning pixels to a set of learnable prototypes, which can adapt to each sample by attentive pooling over the feature and form new slots. Based on the learned data-dependent slots, a contrastive objective is employed for representation learning, which enhances the discriminability of features, and conversely facilitates grouping semantically coherent pixels together. Compared with previous efforts, by simultaneously optimizing the two coupled objectives of semantic grouping and contrastive learning, our approach bypasses the disadvantages of hand-crafted priors and is able to learn object/group-level representations from scene-centric images. Experiments show our approach effectively decomposes complex scenes into semantic groups for feature learning and significantly benefits downstream tasks, including object detection, instance segmentation, and semantic segmentation.

## [Gradient Descent Is Optimal Under Lower Restricted Secant Inequality And Upper Error Bound](#)

- Charles Guille-Escuret · Adam Ibrahim · Baptiste Goujaud · Ioannis Mitliagkas
- abstract@[open-review](#): The study of first-order optimization is sensitive to the assumptions made on the objective functions. These assumptions induce complexity classes which play a key role in worst-case analysis, including the fundamental concept of algorithm optimality. Recent work argues that strong convexity and smoothness are popular assumptions in literature and lead to a pathological definition of the condition number. Motivated by this result, we focus on the class of functions satisfying a lower restricted secant inequality and an upper error bound. On top of being robust to the aforementioned pathological behavior and including some non-convex functions, this pair of conditions displays interesting geometrical properties. In particular, the necessary and sufficient conditions to interpolate a set of points and their gradients within the class can be separated into simple conditions on each sampled gradient. This allows the performance estimation problem (PEP) to be solved analytically, leading to a lower bound on the convergence rate that proves gradient descent to be exactly optimal on this class of functions among all first-order algorithms.

## [Which Explanation Should I Choose? A Function Approximation Perspective to Characterizing Post hoc Explanations](#)

- Tessa Han · Suraj Srinivas · Himabindu Lakkaraju
- abstract@[open-review](#): Despite the plethora of post hoc model explanation methods, the basic properties and behavior of these methods and the conditions under which each one is effective are not well understood. In this work, we bridge these gaps and address a fundamental question: Which explanation method should one use in a given situation? To this end, we adopt a function approximation perspective and formalize the local function approximation (LFA) framework. We show that popular explanation methods are instances of this framework, performing function approximations of the underlying model in different neighborhoods using different loss functions. We introduce a no free lunch theorem for explanation methods which demonstrates that no single method can perform optimally across all neighbourhoods and calls for choosing among methods. To choose among methods, we set forth a guiding principle based on the function approximation perspective, considering a method to be effective if it recovers the underlying model when the model is a member of the explanation function class. Then, we analyze the conditions under which popular explanation methods are effective and provide recommendations for choosing among explanation methods and creating new ones. Lastly, we empirically validate our theoretical results using various real world datasets, model classes, and prediction tasks. By providing a principled mathematical framework which unifies diverse explanation methods, our work characterizes the behaviour of these methods and their relation to one another, guides the choice of explanation methods, and paves the way for the creation of new ones.

## [Alignment-guided Temporal Attention for Video Action Recognition](#)

- Yizhou Zhao · Zhenyang Li · Xun Guo · Yan Lu
- abstract@[open-review](#): Temporal modeling is crucial for various video learning tasks. Most recent approaches employ either factorized (2D+1D) or joint (3D) spatial-temporal operations to extract temporal contexts from the input frames. While the former is more efficient in computation, the latter often obtains better performance. In this paper, we attribute this to a dilemma between the sufficiency and the efficiency of interactions among various positions in different frames. These interactions affect the extraction of task-relevant information shared among frames. To resolve this issue, we prove that frame-by-frame alignments have the potential to increase the mutual information between frame representations, thereby including more task-relevant information to boost effectiveness. Then we propose Alignment-guided Temporal Attention (ATA) to extend 1-dimensional temporal attention with parameter-free patch-level alignments between neighboring frames. It can act as a general plug-in for image backbones to conduct the action recognition task without any model-specific design. Extensive experiments on multiple benchmarks demonstrate the superiority and generality of our module.

## [Stochastic Multiple Target Sampling Gradient Descent](#)

- Hoang Phan · Ngoc Tran · Trung Le · Toan Tran · Nhat Ho · Dinh Phung
- abstract@[open-review](#): Sampling from an unnormalized target distribution is an essential problem with many applications in probabilistic inference. Stein Variational Gradient Descent (SVGD) has been shown to be a powerful method that iteratively updates a set of particles to approximate the distribution of interest. Furthermore, when analysing its asymptotic properties, SVGD reduces exactly to a single-objective optimization problem and can be viewed as a probabilistic version of this single-objective optimization problem. A natural question then arises: ``Can we derive a probabilistic version of the multi-objective optimization?". To answer this question, we propose Stochastic Multiple Target Sampling Gradient Descent (MT-SGD), enabling us to sample from multiple unnormalized target distributions. Specifically, our MT-SGD conducts a flow of intermediate distributions gradually orienting to multiple target distributions, which allows the sampled particles to move to the joint high-likelihood region of the target distributions. Interestingly, the asymptotic

analysis shows that our approach reduces exactly to the multiple-gradient descent algorithm for multi-objective optimization, as expected. Finally, we conduct comprehensive experiments to demonstrate the merit of our approach to multi-task learning.

## [Neural Networks with Hadamard Product: Separation on Extrapolation and Spectral Bias](#)

- Yongtao Wu · Zhenyu Zhu · Fanghui Liu · Grigoris Chrysos · Volkan Cevher
- abstract@[open-review](#): Neural tangent kernel (NTK) is a powerful tool to analyze training dynamics of neural networks and their generalization bounds. The study on NTK has been devoted to typical neural network architectures, but is incomplete for neural networks with Hadamard products (NNs-Hp), e.g., StyleGAN and polynomial neural networks. In this work, we derive the finite-width NTK formulation for a special class of NNs-Hp, i.e., polynomial neural networks. We prove their equivalence to the kernel regression predictor with the associated NTK, which expands the application scope of NTK. Based on our results, we elucidate the separation of PNNs over standard neural networks with respect to extrapolation and spectral bias. Our two key insights are that when compared to standard neural networks, PNNs are able to fit more complicated functions in the extrapolation regime and admit a slower eigenvalue decay of the respective NTK. Besides, our theoretical results can be extended to other types of NNs-Hp, which expand the scope of our work. Our empirical results validate the separations in broader classes of NNs-Hp, which provide a good justification for a deeper understanding of neural architectures.

## [Orthogonal Transformer: An Efficient Vision Transformer Backbone with Token Orthogonalization](#)

- Huaibo Huang · Xiaoqiang Zhou · Ran He
- abstract@[open-review](#): We present a general vision transformer backbone, called as Orthogonal Transformer, in pursuit of both efficiency and effectiveness. A major challenge for vision transformer is that self-attention, as the key element in capturing long-range dependency, is very computationally expensive for dense prediction tasks (e.g., object detection). Coarse global self-attention and local self-attention are then designed to reduce the cost, but they suffer from either neglecting local correlations or hurting global modeling. We present an orthogonal self-attention mechanism to alleviate these issues. Specifically, self-attention is computed in the orthogonal space that is reversible to the spatial domain but has much lower resolution. The capabilities of learning global dependency and exploring local correlations are maintained because every orthogonal token in self-attention can attend to the entire visual tokens. Remarkably, orthogonality is realized by constructing an endogenously orthogonal matrix that is friendly to neural networks and can be optimized as arbitrary orthogonal matrices. We also introduce Positional MLP to incorporate position information for arbitrary input resolutions as well as enhance the capacity of MLP. Finally, we develop a hierarchical architecture for Orthogonal Transformer. Extensive experiments demonstrate its strong performance on a broad range of vision tasks, including image classification, object detection, instance segmentation and semantic segmentation.

## [On-Device Training Under 256KB Memory](#)

- Ji Lin · Ligeng Zhu · Wei-Ming Chen · Wei-Chen Wang · Chuang Gan · Song Han
- abstract@[open-review](#): On-device training enables the model to adapt to new data collected from the sensors. However, the training memory consumption is prohibitive for IoT devices that have tiny memory resources. We propose an algorithm-system co-design framework to make training neural networks possible with only 256KB of memory. On-device training faces two unique challenges: the quantized graphs of neural networks are hard to optimize due to mixed bit-precision and the lack of normalization; the limited hardware resource (memory and computation) does not allow full backward computation. To cope with the optimization difficulty, we propose Quantization-Aware Scaling to calibrate the gradient scales and stabilize quantized training. To reduce the memory footprint, we propose Sparse Update to skip the gradient computation of less important layers and sub-tensors. The algorithm innovation is implemented by a lightweight training system, Tiny Training Engine, which prunes the backward computation graph to support sparse updates and offload the runtime auto-differentiation to compile time. Our framework is the first practical solution for on-device transfer learning of visual recognition on tiny IoT devices (e.g., a microcontroller with only 256KB SRAM), using less than 1/100 of the memory of existing frameworks and matching the accuracy of cloud training+edge deployment for the tinyML application VWW. Our study suggests that tiny IoT devices can not only perform inference but also continuously adapt to new data for lifelong learning.

## [Evaluating Graph Generative Models with Contrastively Learned Features](#)

- Hamed Shirzad · Kaveh Hassani · Danica J. Sutherland
- abstract@[open-review](#): A wide range of models have been proposed for Graph Generative Models, necessitating effective methods to evaluate their quality. So far, most techniques use either traditional metrics based on subgraph counting, or the representations of randomly initialized Graph Neural Networks (GNNs). We propose using representations from contrastively trained GNNs, rather than random GNNs, and show this gives more reliable evaluation metrics. Neither traditional approaches nor GNN-based approaches dominate the other, however: we give examples of graphs that each approach is unable to distinguish. We demonstrate that Graph Substructure Networks (GSNs), which in a way combine both approaches, are better at distinguishing the distances between graph datasets.

## [Fast Distance Oracles for Any Symmetric Norm](#)

- Yichuan Deng · Zhao Song · OMRI WEINSTEIN · Ruizhe Zhang
- abstract@[open-review](#): In the `Distance Oracle` problem, the goal is to preprocess  $n$  vectors  $x_1, x_2, \dots, x_n$  in a  $d$ -dimensional normed space  $(\mathbb{X}^d, \| \cdot \|)$  into a cheap data structure, so that given a query vector  $q \in \mathbb{X}^d$ , all distances  $\| q - x_i \|$  to the data points  $x_i$  ( $i \in [n]$ ) can be quickly approximated (faster than the trivial  $\sim n^2$  query time). This primitive is a basic subroutine in machine learning, data mining and similarity search applications. In the case of  $\ell_p$  norms, the problem is well understood, and optimal data structures are known for most values of  $p$ . Our main contribution is a fast  $(1+\epsilon)$  distance oracle for any symmetric norm  $\|\cdot\|$ . This class includes  $\ell_p$  norms and Orlicz norms as special cases, as well as other norms used in practice, e.g. top- $k$  norms, max-mixture and sum-mixture of  $\ell_p$  norms, small-support norms and the box-norm. We propose a novel data structure with  $\tilde{O}(n(d + \mathsf{mmc}(l)^2))$  preprocessing time and space, and  $\tilde{t}_q = \tilde{O}(d + n \cdot \mathsf{mmc}(l)^2)$  query time, where  $\mathsf{mmc}(l)$  is a complexity-measure (modulus) of the symmetric norm under consideration. When  $l = \ell_p$ , this runtime matches the aforementioned state-of-art oracles.

## [Regularized Molecular Conformation Fields](#)

- Lihao Wang · Yi Zhou · Yiqun Wang · Xiaoqing Zheng · Xuanjing Huang · Hao Zhou
- abstract@[open-review](#): Predicting energetically favorable 3-dimensional conformations of organic molecules from molecular graph plays a fundamental role in computer-aided drug discovery research. However, effectively exploring the high-dimensional conformation space to identify (meta) stable conformers is anything but trivial. In this work, we introduce RMCF, a novel framework to generate a diverse set of low-energy molecular conformations through sampling from a regularized molecular conformation field. We develop a data-driven molecular segmentation algorithm to automatically partition each molecule into several structural building blocks to reduce the modeling degrees of freedom. Then, we employ a Markov Random Field to learn the joint probability distribution of fragment configurations and inter-fragment dihedral angles, which enables us to sample from different low-energy regions of a conformation space. Our model constantly outperforms state-of-the-art models for the conformation generation task on the GEOM-Drugs dataset. We attribute the success of RMCF to modeling in a regularized feature space and learning a global fragment configuration distribution for effective sampling. The proposed method could be generalized to deal with larger biomolecular systems.

## [Learning single-index models with shallow neural networks](#)

- Alberto Bietti · Joan Bruna · Clayton Sanford · Min Jae Song
- abstract@[open-review](#): Single-index models are a class of functions given by an unknown univariate ``link'' function applied to an unknown one-dimensional projection of the input. These models are particularly relevant in high dimension, when the data might present low-dimensional structure that learning algorithms should adapt to. While several statistical aspects of this model, such as the sample complexity of recovering the relevant (one-dimensional) subspace, are well-understood, they rely on tailored algorithms that exploit the specific structure of the target function. In this work, we introduce a natural class of shallow neural networks and study its ability to learn single-index models via \textit{gradient descent}. More precisely, we consider shallow networks in which the first layer weights are tied, to mirror the single-index model structure, and biases of the neurons are frozen at random initialization. We show that the corresponding optimization landscape is benign, which in turn leads to generalization guarantees that match the optimal sample complexity of dedicated semi-parametric methods.

## [Generating Training Data with Language Models: Towards Zero-Shot Language Understanding](#)

- Yu Meng · Jiaxin Huang · Yu Zhang · Jiawei Han
- abstract@[open-review](#): Pretrained language models (PLMs) have demonstrated remarkable performance in various natural language processing tasks: Unidirectional PLMs (e.g., GPT) are well known for their superior text generation capabilities; bidirectional PLMs (e.g., BERT) have been the prominent choice for natural language understanding (NLU) tasks. While both types of models have achieved promising few-shot learning performance, their potential for zero-shot learning has been underexplored. In this paper, we present a simple approach that uses both types of PLMs for fully zero-shot learning of NLU tasks without requiring any task-specific data: A unidirectional PLM generates class-conditioned texts guided by prompts, which are used as the training data for fine-tuning a bidirectional PLM. With quality training data selected based on the generation probability and regularization techniques (label smoothing and temporal ensembling) applied to the fine-tuning stage for better generalization and stability, our approach demonstrates strong performance across seven classification tasks of the GLUE benchmark (e.g., 72.3/73.8 on MNLI-m/mm and 92.8 on SST-2), significantly outperforming zero-shot prompting methods and achieving even comparable results to strong few-shot approaches using 32 training samples per class.

## [SageMix: Saliency-Guided Mixup for Point Clouds](#)

- Sanghyeon Lee · Minkyu Jeon · Injae Kim · Yunyang Xiong · Hyunwoo Kim
- abstract@[open-review](#): Data augmentation is key to improving the generalization ability of deep learning models. Mixup is a simple and widely-used data augmentation technique that has proven effective in alleviating the problems of overfitting and data scarcity. Also, recent studies of saliency-aware Mixup in the image domain show that preserving discriminative parts is beneficial to improving the generalization performance. However, these Mixup-based data augmentations are underexplored in 3D vision, especially in point clouds. In this paper, we propose SageMix, a saliency-guided Mixup for point clouds to preserve salient local structures. Specifically, we extract salient regions from two point clouds and smoothly combine them into one continuous shape. With a simple sequential sampling by re-weighted saliency scores, SageMix preserves the local structure of salient regions. Extensive experiments demonstrate that the proposed method consistently outperforms existing Mixup methods in various benchmark point cloud datasets. With PointNet++, our method achieves an accuracy gain of 2.6% and 4.0% over standard training in ModelNet40 and ScanObjectNN, respectively. In addition to generalization performance, SageMix improves robustness and uncertainty calibration. Moreover, when adopting our method to various tasks including part segmentation and standard image classification, our method achieves competitive performance.

## [Efficient \\$\Phi\\$-Regret Minimization in Extensive-Form Games via Online Mirror Descent](#)

- Yu Bai · Chi Jin · Song Mei · Ziang Song · Tiancheng Yu
- abstract@[open-review](#): A conceptually appealing approach for learning Extensive-Form Games (EFGs) is to convert them to Normal-Form Games (NFGs). This approach enables us to directly translate state-of-the-art techniques and analyses in NFGs to learning EFGs, but typically suffers from computational intractability due to the exponential blow-up of the game size introduced by the conversion. In this paper, we address this problem in natural and important setups for the \emph{\$\Phi\$-Hedge} algorithm---A generic algorithm capable of learning a large class of equilibria for NFGs. We show that \$\Phi\$-Hedge can be directly used to learn Nash Equilibria (zero-sum settings), Normal-Form Coarse Correlated Equilibria (NFCCE), and Extensive-Form Correlated Equilibria (EFCE) in EFGs. We prove that, in those settings, the \emph{\$\Phi\$-Hedge} algorithms are equivalent to standard Online Mirror Descent (OMD) algorithms for EFGs with suitable dilated regularizers, and run in polynomial time. This new connection further allows us to design and analyze a new class of OMD algorithms based on modifying its log-partition function. In particular, we design an improved algorithm with balancing techniques that achieves a sharp \$\widetilde{\mathcal{O}}(\sqrt{XAT})\$ EFCE-regret under bandit-feedback in an EFG with \$X\$ information sets, \$A\$ actions, and \$T\$ episodes. To our best knowledge, this is the first such rate and matches the information-theoretic lower bound.

## [Two-layer neural network on infinite dimensional data: global optimization guarantee in the mean-field regime](#)

- Naoki Nishikawa · Taiji Suzuki · Atsushi Nitanda · Denny Wu
- abstract@[open-review](#): Analysis of neural network optimization in the mean-field regime is important as the setting allows for feature learning. Existing theory has been developed mainly for neural networks in finite dimensions, i.e., each neuron has a finite-dimensional parameter. However, the setting of infinite-dimensional input naturally arises in machine learning problems such as nonparametric functional data analysis and graph classification. In this paper, we develop a new mean-field analysis of two-layer neural network in an infinite-dimensional parameter space. We first give a generalization error bound, which shows that the regularized empirical risk minimizer properly generalizes when the data size is sufficiently large, despite the neurons being infinite-dimensional. Next, we present two gradient-based optimization algorithms for infinite-dimensional mean-field networks, by extending the recently developed particle optimization framework to the infinite-dimensional setting. We show that the proposed algorithms converge to the (regularized) global optimal solution, and moreover, their rates of convergence are of polynomial order in the online setting and exponential order in the finite sample setting, respectively. To our knowledge this is the first quantitative global optimization guarantee of neural network on infinite-dimensional input and in the presence of feature learning.

## [Escaping Saddle Points with Bias-Variance Reduced Local Perturbed SGD for Communication Efficient Nonconvex Distributed Learning](#)

- Tomoya Murata · Taiji Suzuki
- abstract@[open-review](#): In recent centralized nonconvex distributed learning and federated learning, local methods are one of the promising approaches to reduce communication time. However, existing work has mainly focused on studying first-order optimality guarantees. On the other side, second-order optimality guaranteed algorithms, i.e., algorithms escaping saddle points, have been extensively studied in the non-distributed optimization literature. In this paper, we study a new local algorithm called Bias-Variance Reduced Local Perturbed SGD (BVR-L-PSGD), that combines the existing bias-variance reduced gradient estimator with parameter perturbation to find second-order optimal points in centralized nonconvex distributed optimization. BVR-L-PSGD enjoys second-order optimality with nearly the same communication complexity as the best known one of BVR-L-SGD to find first-order optimality. Particularly, the communication complexity is better than non-local methods when the local datasets heterogeneity is smaller than the smoothness of the local loss. In an extreme case, the communication complexity approaches to \$\widetilde{\Theta}(1)\$ when the local datasets heterogeneity goes to zero. Numerical results validate our theoretical findings.

## [UniCLIP: Unified Framework for Contrastive Language-Image Pre-training](#)

- Janghyeon Lee · Jongsuk Kim · Hyounguk Shon · Bumsoo Kim · Seung Hwan Kim · Honglak Lee · Junmo Kim
- abstract@[open-review](#): Pre-training vision-language models with contrastive objectives has shown promising results that are both scalable to large uncurated datasets and transferable to many downstream applications. Following works have targeted to improve data efficiency by adding self-supervision terms. However, as these works define inter-domain (image-text) contrastive loss and intra-domain (image-image) contrastive loss in individual spaces, many feasible combinations of supervision are overlooked. To overcome this issue, we propose UniCLIP, a Unified framework for Contrastive Language-Image Pre-training. UniCLIP integrates the contrastive loss of both inter-domain pairs and intra-domain pairs into a single universal space. The discrepancies that occur when integrating contrastive loss between different domains are resolved by the three key components of UniCLIP: (1) augmentation-aware feature embedding, (2) MP-NCE loss, and (3) domain dependent similarity measure. UniCLIP outperforms previous vision-language pre-training methods throughout various single- and multi-modality downstream tasks. In our experiments, we show that each component that comprises UniCLIP contributes well to the final performance.

## [Deep Model Reassembly](#)

- Xingyi Yang · Daquan Zhou · Songhua Liu · Jingwen Ye · Xinchao Wang
- abstract@[open-review](#): In this paper, we explore a novel knowledge-transfer task, termed as Deep Model Reassembly (DeRy), for general-purpose model reuse. Given a collection of heterogeneous models pre-trained from distinct sources and with diverse architectures, the goal of DeRy, as its name implies, is to first dissect each model into distinctive building blocks, and then selectively reassemble the derived blocks to produce customized networks under both the hardware resource and performance constraints. Such ambitious nature of DeRy inevitably imposes significant challenges, including, in the first place, the feasibility of its solution. We strive to showcase that, through a dedicated paradigm proposed in this paper, DeRy can be made not only possibly but practically efficiently. Specifically, we conduct the partitions of all pre-trained networks jointly via a cover set optimization, and derive a number of equivalence sets, within each of which the network blocks are treated as functionally equivalent and hence interchangeable. The equivalence sets learned in this way, in turn, enable picking and assembling blocks to customize networks subject to certain constraints, which is achieved via solving an integer program backed up with a training-free proxy to estimate the task performance. The reassembled models give rise to gratifying performances with the user-specified constraints satisfied. We demonstrate that on ImageNet, the best reassemble model achieves 78.6% top-1 accuracy without fine-tuning, which could be further elevated to 83.2% with end-to-end fine-tuning. Our code will be made publicly available.

## [Learning to Configure Computer Networks with Neural Algorithmic Reasoning](#)

- Luca Beurer-Kellner · Martin Vechev · Laurent Vanbever · Petar Veličković
- abstract@[open-review](#): We present a new method for scaling automatic configuration of computer networks. The key idea is to relax the computationally hard search problem of finding a configuration that satisfies a given specification into an approximate objective amenable to learning-based techniques. Based on this idea, we train a neural algorithmic model which learns to generate configurations likely to (fully or partially) satisfy a given specification under existing routing protocols. By relaxing the rigid satisfaction guarantees, our approach (i) enables greater flexibility: it is protocol-agnostic, enables cross-protocol reasoning, and does not depend on hardcoded rules; and (ii) finds configurations for much larger computer networks than previously possible. Our learned synthesizer is up to 490x faster than state-of-the-art SMT-based methods, while producing configurations which on average satisfy more than 93% of the provided requirements.

## [Identifiability of deep generative models under mixture priors without auxiliary information](#)

- Bohdan Kivva · GOUTHAM RAJENDRAN · Pradeep Ravikumar · Bryon Aragam
- abstract@[open-review](#): We prove identifiability of a broad class of deep latent variable models that (a) have universal approximation capabilities and (b) are the decoders of variational autoencoders that are commonly used in practice. Unlike existing work, our analysis does not require weak supervision, auxiliary information, or conditioning in the latent space. Recently, there has been a surge of works studying identifiability of such models. In these works, the main assumption is that along with the data, an auxiliary variable \$u\$ (also known as side information) is observed as well. At the same time, several works have empirically observed that this doesn't seem to be necessary in practice. In this work, we explain this behavior by showing that for a broad class of generative (i.e. unsupervised) models with universal approximation capabilities, the side information \$u\$ is not necessary: We prove identifiability of the entire generative model where we do not observe \$u\$ and only observe the data \$x\$. The models we consider are tightly connected with autoencoder architectures used in practice that leverage mixture priors in the latent space and ReLU/leaky-ReLU activations in the encoder. Our main result is an identifiability hierarchy that significantly generalizes previous work and exposes how different assumptions lead to different ``strengths'' of identifiability. For example, our weakest result establishes (unsupervised) identifiability up to an affine transformation, which already improves existing work. It's well known that these models have universal approximation capabilities and moreover, they have been extensively used in practice to learn representations of data.

## [On A Mallows-type Model For \(Ranked\) Choices](#)

- Yifan Feng · Yuxuan Tang
- abstract@[open-review](#): We consider a preference learning setting where every participant chooses an ordered list of \$k\$ most preferred items among a displayed set of candidates. (The set can be different for every participant.) We identify a distance-based ranking model for the population's preferences and their (ranked) choice behavior. The ranking model resembles the Mallows model but uses a new distance function called Reverse Major Index (RMJ). We find that despite the need to sum over all permutations, the RMJ-based ranking distribution aggregates into (ranked) choice probabilities with simple closed-form expression. We develop effective methods to estimate the model parameters and showcase their generalization power using real data, especially when there is a limited variety of display sets.

## [Parameters or Privacy: A Provable Tradeoff Between Overparameterization and Membership Inference](#)

- Jasper Tan · Blake Mason · Hamid Javadi · Richard Baraniuk
- abstract@[open-review](#): A surprising phenomenon in modern machine learning is the ability of a highly {\em overparameterized} model to generalize well (small error on the test data) even when it is trained to memorize the training data (zero error on the training data). This has led to an arms race towards increasingly overparameterized models (c.f., deep learning). In this paper, we study an underexplored hidden cost of overparameterization: the fact that overparameterized models are more vulnerable to {\em privacy attacks}, in particular the {\em membership inference} attack that predicts the (potentially sensitive) examples used to train a model. We significantly extend the relatively few empirical results on this problem by theoretically proving for an overparameterized linear regression model with Gaussian data that the membership inference vulnerability increases with the number of parameters. Moreover, a range of empirical studies indicates that more complex, nonlinear models exhibit the same behavior. Finally, we study different methods for mitigating such attacks in the overparameterized regime, such as noise addition and regularization, and conclude that simply reducing the parameters of an overparameterized model is an effective strategy to protect it from membership inference without greatly decreasing its generalization error.

## [NeurOLight: A Physics-Agnostic Neural Operator Enabling Parametric Photonic Device Simulation](#)

- Jiaqi Gu · Zhengqi Gao · Chenghao Feng · Hanqing Zhu · Ray Chen · Duane Boning · David Pan

- abstract@[open-review](#): Optical computing has become emerging technology in next-generation efficient artificial intelligence (AI) due to its ultra-high speed and efficiency. Electromagnetic field simulation is critical to the design, optimization, and validation of photonic devices and circuits. However, costly numerical simulation significantly hinders the scalability and turn-around time in the photonic circuit design loop. Recently, physics-informed neural networks were proposed to predict the optical field solution of a single instance of a partial differential equation (PDE) with predefined parameters. Their complicated PDE formulation and lack of efficient parametrization mechanism limit their flexibility and generalization in practical simulation scenarios. In this work, for the first time, a physics-agnostic neural operator-based framework, dubbed NeurOLight, is proposed to learn a family of frequency-domain Maxwell PDEs for ultra-fast parametric photonic device simulation. Specifically, we discretize different devices into a unified domain, represent parametric PDEs with a compact wave prior, and encode the incident light via masked source modeling. We design our model to have parameter-efficient cross-shaped NeurOLight blocks and adopt superposition-based augmentation for data-efficient learning. With those synergistic approaches, NeurOLight demonstrates 2-orders-of-magnitude faster simulation speed than numerical solvers and outperforms prior NN-based models by ~54% lower prediction error using ~44% fewer parameters.

## [Spherization Layer: Representation Using Only Angles](#)

- Hoyong Kim Â· kangil kim
- abstract@[open-review](#): In neural network literature, angular similarity between feature vectors is frequently used for interpreting or re-using learned representations. However, the inner product in neural networks partially disperses information over the scales and angles of the involved input vectors and weight vectors. Therefore, when using only angular similarity on representations trained with the inner product, information loss occurs in downstream methods, which limits their performance. In this study, we proposed spherization layer to represent all information on angular similarity. The layer 1) maps the pre-activations of input vectors into the specific range of angles, 2) converts the angular coordinates of the vectors to Cartesian coordinates with an additional dimension, and 3) trains decision boundaries from hyperplanes, without bias parameters, passing through the origin. This approach guarantees that representation learning always occurs on the hyperspherical surface without the loss of any information unlike other projection-based methods. Furthermore, this method can be applied to any network by replacing an existing layer. We validate the functional correctness of the proposed method in a toy task, retention ability in well-known image classification tasks, and effectiveness in few-shot learning, word analogy test, and hyperspherical learning.

## [A Deep Learning Dataloader with Shared Data Preparation](#)

- jian xie Â· Jingwei Xu Â· Guochang Wang Â· Yuan Yao Â· Zenan Li Â· Chun Cao Â· Hanghang Tong
- abstract@[open-review](#): Executing a family of Deep Neural Networks (DNNs) training jobs on the same or similar datasets in parallel is typical in current deep learning scenarios. It is time-consuming and resource-intensive because each job repetitively prepares (i.e., loads and preprocesses) the data independently, causing redundant consumption of I/O and computations. Although the page cache or a centralized cache component can alleviate the redundancies by reusing the data prep work, each job's data sampled uniformly at random presents a low sampling locality in the shared dataset that causes the heavy cache thrashing. Prior work tries to solve the problem by enforcing all training jobs iterating over the dataset in the same order and requesting each data in lockstep, leading to strong constraints: all jobs must have the same dataset and run simultaneously. In this paper, we propose a dependent sampling algorithm (DSA) and domain-specific cache policy to relax the constraints. Besides, a novel tree data structure is designed to efficiently implement DSA. Based on the proposed technologies, we implemented a prototype system, named Joader, which can share data prep work as long as the datasets share partially. We evaluate the proposed Joader in practical scenarios, showing a greater versatility and superiority over training speed improvement (up to 500% in ResNet18).

## [Are You Stealing My Model? Sample Correlation for Fingerprinting Deep Neural Networks](#)

- Jiyang Guan Â· Jian Liang Â· Ran He
- abstract@[open-review](#): An off-the-shelf model as a commercial service could be stolen by model stealing attacks, posing great threats to the rights of the model owner. Model fingerprinting aims to verify whether a suspect model is stolen from the victim model, which gains more and more attention nowadays. Previous methods always leverage the transferable adversarial examples as the model fingerprint, which is sensitive to adversarial defense or transfer learning scenarios. To address this issue, we consider the pairwise relationship between samples instead and propose a novel yet simple model stealing detection method based on SAmple Correlation (SAC). Specifically, we present SAC-w that selects wrongly classified normal samples as model inputs and calculates the mean correlation among their model outputs. To reduce the training time, we further develop SAC-m that selects mixed samples via CutMix as model inputs, without the need for training the surrogate models or generating adversarial examples. Extensive results validate that both SAC-w and SAC-m successfully defend against various model stealing attacks, even including adversarial training or transfer learning, and detect the stolen models with the best performance in terms of AUC across different datasets and model architectures. Code is attached in the supplementary.

## [Large-scale Optimization of Partial AUC in a Range of False Positive Rates](#)

- Yao Yao Â· Qihang Lin Â· Tianbao Yang
- abstract@[open-review](#): The area under the ROC curve (AUC) is one of the most widely used performance measures for classification models in machine learning. However, it summarizes the true positive rates (TPRs) over all false positive rates (FPRs) in the ROC space, which may include the FPRs with no practical relevance in some applications. The partial AUC, as a generalization of the AUC, summarizes only the TPRs over a specific range of the FPRs and is thus a more suitable performance measure in many real-world situations. Although partial AUC optimization in a range of FPRs had been studied, existing algorithms are not scalable to big data and not applicable to deep learning. To address this challenge, we cast the problem into a non-smooth difference-of-convex (DC) program for any smooth predictive functions (e.g., deep neural networks), which allowed us to develop an efficient approximated gradient descent method based on the Moreau envelope smoothing technique, inspired by recent advances in non-smooth DC optimization. To increase the efficiency of large data processing, we used an efficient stochastic block coordinate update in our algorithm. Our proposed algorithm can also be used to minimize the sum of ranked range loss, which also lacks efficient solvers. We established a complexity of  $\tilde{O}(1/\epsilon^6)$  for finding a nearly  $\epsilon$ -critical solution. Finally, we numerically demonstrated the effectiveness of our proposed algorithms in training both linear models and deep neural networks for partial AUC maximization and sum of ranked range loss minimization.

## [PointTAD: Multi-Label Temporal Action Detection with Learnable Query Points](#)

- Jing Tan Â· Xiaotong Zhao Â· Xintian Shi Â· Bin Kang Â· Limin Wang
- abstract@[open-review](#): Traditional temporal action detection (TAD) usually handles untrimmed videos with small number of action instances from a single label (e.g., ActivityNet, THUMOS). However, this setting might be unrealistic as different classes of actions often co-occur in practice. In this paper, we focus on the complex multi-label temporal action detection that aims to localize all action instances from a multi-label untrimmed video. Multi-label TAD is more challenging as it requires for fine-grained class discrimination within a single video and dedicated module to precisely localize the co-occurring instances. Specifically, we extend the sparse query-based detection paradigm from the traditional TAD and propose the multi-label TAD framework of PointTAD. Existing query-based action detectors employs a segment to represent an action instance, which is insufficient to handle the concurrent instances and their richer relations. To mitigate this issue, our PointTAD introduces a small set of learnable query points to represent important frames of each action instance. Our PointTAD provides a flexible mechanism to localize the discriminative frames at boundaries and as well the important frames inside the action. Moreover, we improve action decoding with the Multi-level Interactive Module to integrate action semantics at point-level and instance-level. Finally, our PointTAD employs an end-to-end trainable framework based on RGB input for easy deployment. We evaluate our proposed

method on two popular benchmarks for multi-label TAD. Our model outperforms all previous methods by a large margin under the detection-mAP metric and achieves promising results under the segmentation-mAP metric.

## [Improve Task-Specific Generalization in Few-Shot Learning via Adaptive Vicinal Risk Minimization](#)

- Long-Kai Huang · Ying Wei
- abstract@[open-review](#): Recent years have witnessed the rapid development of meta-learning in improving the meta generalization over tasks in few-shot learning. However, the task-specific level generalization is overlooked in most algorithms. For a novel few-shot learning task where the empirical distribution likely deviates from the true distribution, the model obtained via minimizing the empirical loss can hardly generalize to unseen data. A viable solution to improving the generalization comes as a more accurate approximation of the true distribution; that is, admitting a Gaussian-like vicinal distribution for each of the limited training samples. Thereupon we derive the resulting vicinal loss function over vicinities of all training samples and minimize it instead of the conventional empirical loss over training samples only, favorably free from the exhaustive sampling of all vicinal samples. It remains challenging to obtain the statistical parameters of the vicinal distribution for each sample. To tackle this challenge, we further propose to estimate the statistical parameters as the weighted mean and variance of a set of unlabeled data it passed by a random walk starting from training samples. To verify the performance of the proposed method, we conduct experiments on four standard few-shot learning benchmarks and consolidate the superiority of the proposed method over state-of-the-art few-shot learning baselines.

## [Adaptive Sampling for Discovery](#)

- Ziping Xu · Eunjae Shim · Ambuj Tewari · Paul Zimmerman
- abstract@[open-review](#): In this paper, we study a sequential decision-making problem, called Adaptive Sampling for Discovery (ASD). Starting with a large unlabeled dataset, algorithms for ASD adaptively label the points with the goal to maximize the sum of responses. This problem has wide applications to real-world discovery problems, for example drug discovery with the help of machine learning models. ASD algorithms face the well-known exploration-exploitation dilemma. The algorithm needs to choose points that yield information to improve model estimates but it also needs to exploit the model. We rigorously formulate the problem and propose a general information-directed sampling (IDS) algorithm. We provide theoretical guarantees for the performance of IDS in linear, graph and low-rank models. The benefits of IDS are shown in both simulation experiments and real-data experiments for discovering chemical reaction conditions.

## [Learning Representations via a Robust Behavioral Metric for Deep Reinforcement Learning](#)

- Jianda Chen · Sinno Pan
- abstract@[open-review](#): Learning an informative representation with behavioral metrics is able to accelerate deep reinforcement learning process. Two key research issues on behavioral metric-based representation learning are how to relax the computation of a specific behavioral metric, which is difficult or even intractable to compute, and how to approximate the relaxed metric by learning an embedding space for states. In this paper, we analyze the potential relaxation and/or approximation gaps for existing behavioral metric-based representation learning methods. Based on the analysis, we propose a new behavioral distance, the RAP distance, and develop a practical representation learning algorithm on top of it. We provide theoretical analysis on the proposed algorithm. We conduct extensive experiments on DeepMind Control Suite with distraction, Robosuite, and autonomous driving simulator CARLA to demonstrate new state-of-the-art results.

## [Learning from Small Samples: Transformation-Invariant SVMs with Composition and Locality at Multiple Scales](#)

- Tao Liu · P. R. Kumar · Ruida Zhou · Xi Liu
- abstract@[open-review](#): Motivated by the problem of learning with small sample sizes, this paper shows how to incorporate into support-vector machines (SVMs) those properties that have made convolutional neural networks (CNNs) successful. Particularly important is the ability to incorporate domain knowledge of invariances, e.g., translational invariance of images. Kernels based on the maximum similarity over a group of transformations are not generally positive definite. Perhaps it is for this reason that they have not been studied theoretically. We address this lacuna and show that positive definiteness indeed holds with high probability for kernels based on the maximum similarity in the small training sample set regime of interest, and that they do yield the best results in that regime. We also show how additional properties such as their ability to incorporate local features at multiple spatial scales, e.g., as done in CNNs through max pooling, and to provide the benefits of composition through the architecture of multiple layers, can also be embedded into SVMs. We verify through experiments on widely available image sets that the resulting SVMs do provide superior accuracy in comparison to well-established deep neural network benchmarks for small sample sizes.

## [DMAP: a Distributed Morphological Attention Policy for learning to locomote with a changing body](#)

- Alberto Silvio Chiappa · Alessandro Marin Vargas · Alexander Mathis
- abstract@[open-review](#): Reinforcement learning typically seeks to learn control policies in stable environments. Yet, real world scenarios require continuous adaptation. In particular, learning to locomote when the length and the thickness of different body parts vary is challenging, as the policy is required to adapt to the current configuration to successfully balance and advance the agent. We study this problem in four classical continuous control environments, augmented with morphological perturbations. We show that a control policy based on the proprioceptive state performs poorly with highly variable body configurations, while an (oracle) agent with access to a learned encoding of the perturbation performs significantly better. We introduce DMAP, a biologically-inspired, attention-based policy network architecture. It combines a distributed policy, with individual controllers for each joint, and an attention mechanism, to dynamically gate sensory information from different body parts. DMAP can be trained end-to-end in all the considered environments and perturbation intensities, overall matching or surpassing the performance of an oracle agent with access to the morphology information. Thus DMAP, implementing principles of control drawn from the biological world, provides a strong inductive bias for learning challenging sensorimotor tasks. Overall, our work corroborates the power of these principles in challenging locomotion tasks.

## [Learning Infinite-Horizon Average-Reward Restless Multi-Action Bandits via Index Awareness](#)

- GUOJUN XIONG · Shufan Wang · Jian Li
- abstract@[open-review](#): We consider the online restless bandits with average-reward and multiple actions, where the state of each arm evolves according to a Markov decision process (MDP), and the reward of pulling an arm depends on both the current state of the corresponding MDP and the action taken. Since finding the optimal control is typically intractable for restless bandits, existing learning algorithms are often computationally expensive or with a regret bound that is exponential in the number of arms and states. In this paper, we advocate index-aware reinforcement learning (RL) solutions to design RL algorithms operating on a much smaller dimensional subspace by exploiting the inherent structure in restless bandits. Specifically, we first propose novel index policies to address dimensionality concerns, which are provably optimal. We then leverage the indices to develop two low-complexity index-aware RL algorithms, namely, (i) GM-R2MAB, which has access to a generative model; and (ii) UC-R2MAB, which learns the model using an upper confidence style online exploitation method. We prove that both algorithms achieve a sub-linear regret that is only polynomial in the number of arms and states. A key differentiator between our algorithms and existing ones stems from the fact that our RL algorithms contain a novel exploitation that leverages our proposed provably optimal index policies for decision-makings.

## [Pluralistic Image Completion with Probabilistic Mixture-of-Experts](#)

- Xiaobo Xia · Wenhao Yang · Jie Ren · Yewen Li · Yibing Zhan · Bo Han · Tongliang Liu
- abstract@[open-review](#): Pluralistic image completion focuses on generating both visually realistic and diverse results for image completion. Prior methods enjoy the empirical successes of this task. However, their used constraints for pluralistic image completion are argued to be not well interpretable and unsatisfactory from two aspects. First, the constraints for visual reality can be weakly correlated to the objective of image completion or even redundant. Second, the constraints for diversity are designed to be task-agnostic, which causes the constraints to not work well. In this paper, to address the issues, we propose an end-to-end probabilistic method. Specifically, we introduce a unified probabilistic graph model that represents the complex interactions in image completion. The entire procedure of image completion is then mathematically divided into several sub-procedures, which helps efficient enforcement of constraints. The sub-procedure directly related to pluralistic results is identified, where the interaction is established by a Gaussian mixture model (GMM). The inherent parameters of GMM are task-related, which are optimized adaptively during training, while the number of its primitives can control the diversity of results conveniently. We formally establish the effectiveness of our method and demonstrate it with comprehensive experiments.

## [Model-based Safe Deep Reinforcement Learning via a Constrained Proximal Policy Optimization Algorithm](#)

- Ashish K Jayant · Shalabh Bhatnagar
- abstract@[open-review](#): During initial iterations of training in most Reinforcement Learning (RL) algorithms, agents perform a significant number of random exploratory steps, which in the real world limit the practicality of these algorithms as these can lead to potentially dangerous behavior. Hence safe exploration is a critical issue in applying RL algorithms in the real world. This problem has been recently well studied under the Constrained Markov Decision Process (CMDP) Framework, where in addition to single-stage rewards, state transitions receive single-stage costs or penalties as well. The prescribed cost functions are responsible for mapping undesirable behavior at any given time-step to a scalar value. Then we aim to find a feasible policy that maximizes reward returns while constraining the cost returns to be below a prescribed threshold during training as well as deployment. We propose an On-policy Model-based Safe Deep RL algorithm in which we learn the transition dynamics of the environment in an online manner as well as find a feasible optimal policy using Lagrangian Relaxation-based Proximal Policy Optimization. We use an ensemble of neural networks with different initializations to tackle epistemic and aleatoric uncertainty issues faced during environment model learning. We compare our approach with relevant model-free and model-based approaches in Constrained RL using the challenging Safe Reinforcement Learning benchmark - the Open AI Safety Gym. We demonstrate that our algorithm is more sample efficient and results in lower cumulative hazard violations as compared to constrained model-free approaches. Further, our approach shows better reward performance than other constrained model-based approach in the literature.

## [CodeRL: Mastering Code Generation through Pretrained Models and Deep Reinforcement Learning](#)

- Hung Le · Yue Wang · Akhilesh Deepak Gotmare · Silvio Savarese · Steven Chu Hong Hoi
- abstract@[open-review](#): Program synthesis or code generation aims to generate a program that satisfies a problem specification. Recent approaches using large-scale pretrained language models (LMs) have shown promising results, yet they have some critical limitations. In particular, they often follow a standard supervised learning procedure to train a code generation model from natural language problem descriptions and ground-truth programs only. Such paradigm has largely ignored some important but potentially useful signals in the problem specification such as unit tests, either during training or inference stages, which thus results in poor performance when solving complex unseen coding tasks. To address the limitations, we propose ``CodeRL'', a new framework to improve pretrained LMs for program synthesis tasks through deep reinforcement learning (RL). Specifically, during training, we treat the code-generating LM as an actor network, and introduce a critic network that is trained to predict the functional correctness of generated programs and provide dense feedback signals to the actor. During inference, we introduce a new generation procedure with a critical sampling strategy that allows a model to automatically regenerate programs based on feedback from example unit tests and critic scores. For the model backbones, we extended the encoder-decoder architecture of CodeT5 with enhanced learning objectives, larger model sizes, and better pretraining data. Our method not only achieves new SOTA results on the APPS benchmark, but also shows strong zero-shot capability with new SOTA results on the simpler MBPP benchmark.

## [GAGA: Deciphering Age-path of Generalized Self-paced Regularizer](#)

- Xingyu Qu · Diyang Li · Xiaohan Zhao · Bin Gu
- abstract@[open-review](#): Nowadays self-paced learning (SPL) is an important machine learning paradigm that mimics the cognitive process of humans and animals. The SPL regime involves a self-paced regularizer and a gradually increasing age parameter, which plays a key role in SPL but where to optimally terminate this process is still non-trivial to determine. A natural idea is to compute the solution path w.r.t. age parameter (i.e., age-path). However, current age-path algorithms are either limited to the simplest regularizer, or lack solid theoretical understanding as well as computational efficiency. To address this challenge, we propose a novel Generalized Age-path Algorithm (GAGA) for SPL with various self-paced regularizers based on ordinary differential equations (ODEs) and sets control, which can learn the entire solution spectrum w.r.t. a range of age parameters. To the best of our knowledge, GAGA is the first exact path-following algorithm tackling the age-path for general self-paced regularizer. Finally the algorithmic steps of classic SVM and Lasso are described in detail. We demonstrate the performance of GAGA on real-world datasets, and find considerable speedup between our algorithm and competing baselines.

## [Understanding the Evolution of Linear Regions in Deep Reinforcement Learning](#)

- Setareh Cohan · Nam Hee Kim · David Rolnick · Michiel van de Panne
- abstract@[open-review](#): Policies produced by deep reinforcement learning are typically characterised by their learning curves, but they remain poorly understood in many other respects. ReLU-based policies result in a partitioning of the input space into piecewise linear regions. We seek to understand how observed region counts and their densities evolve during deep reinforcement learning using empirical results that span a range of continuous control tasks and policy network dimensions. Intuitively, we may expect that during training, the region density increases in the areas that are frequently visited by the policy, thereby affording fine-grained control. We use recent theoretical and empirical results for the linear regions induced by neural networks in supervised learning settings for grounding and comparison of our results. Empirically, we find that the region density increases only moderately throughout training, as measured along fixed trajectories coming from the final policy. However, the trajectories themselves also increase in length during training, and thus the region densities decrease as seen from the perspective of the current trajectory. We further provide other empirically-driven observations regarding the scaling of visited regions with respect to the number of neurons, changing trajectory lengths during training, and the impact of network depth on region density.

## [Plan To Predict: Learning an Uncertainty-Foreseeing Model For Model-Based Reinforcement Learning](#)

- Zifan Wu · Chao Yu · Chen Chen · Jianye Hao · Hankz Hankui Zhuo
- abstract@[open-review](#): In Model-based Reinforcement Learning (MBRL), model learning is critical since an inaccurate model can bias policy learning via generating misleading samples. However, learning an accurate model can be difficult since the policy is continually updated and the induced distribution over visited states used for model learning shifts accordingly. Prior methods alleviate this issue by quantifying the uncertainty of model-generated samples. However, these methods only quantify the uncertainty passively after the samples were generated, rather than foreseeing the uncertainty before model trajectories fall into those highly uncertain regions. The resulting low-quality samples can induce unstable learning targets and hinder the optimization of the policy. Moreover, while being learned to minimize one-step prediction errors, the model is generally used to predict for multiple steps, leading to a mismatch between the objectives of model learning and model usage. To this end, we propose Plan To Predict (P2P), an MBRL framework that treats the model rollout process as a sequential decision making problem by reversely considering the model as a decision maker and the current policy as the dynamics. In this way, the model can quickly adapt to the current policy and foresee the multi-step future uncertainty when generating

trajectories. Theoretically, we show that the performance of P2P can be guaranteed by approximately optimizing a lower bound of the true environment return. Empirical results demonstrate that P2P achieves state-of-the-art performance on several challenging benchmark tasks.

## [Egocentric Video-Language Pretraining](#)

- Kevin Qinghong Lin · Jinpeng Wang · Mattia Soldan · Michael Wray · Rui Yan · Eric Z. XU · Denial Gao · Rong-Cheng Tu · Wenzhe Zhao · Weijie Kong · Chengfei Cai · WANG HongFa · Dima Damen · Bernard Ghanem · Wei Liu · Mike Zheng Shou
- abstract@[open-review](#): Video-Language Pretraining (VLP), aiming to learn transferable representation to advance a wide range of video-text downstream tasks, has recently received increasing attention. Dominant works that achieve strong performance rely on large-scale, 3rd-person video-text datasets, such as HowTo100M. In this work, we exploit the recently released Ego4D dataset to pioneer Egocentric VLP along three directions. (i) We create EgoClip, a 1st-person video-text pretraining dataset comprising 3.8M clip-text pairs well-chosen from Ego4D, covering a large variety of human daily activities. (ii) We propose a novel pretraining objective, dubbed as EgoNCE, which adapts video-text contrastive learning to egocentric domain by mining egocentric-aware positive and negative samples. (iii) We introduce EgoMCQ, a development benchmark that is close to EgoClip and hence can support effective validation and fast exploration of our design decisions regarding EgoClip and EgoNCE. Furthermore, we demonstrate strong performance on five Egocentric VLP downstream tasks across three egocentric datasets: video-text retrieval on EPIC-KITCHENS-100; action recognition on Charades-Ego; and natural language query, moment query, and object state change classification on Ego4D challenge benchmarks.

## [Exploring the Whole Rashomon Set of Sparse Decision Trees](#)

- Rui Xin · Chudi Zhong · Zhi Chen · Takuya Takagi · Margo Seltzer · Cynthia Rudin
- abstract@[open-review](#): In any given machine learning problem, there may be many models that could explain the data almost equally well. However, most learning algorithms return only one of these models, leaving practitioners with no practical way to explore alternative models that might have desirable properties beyond what could be expressed within a loss function. The Rashomon set is the set of these all almost-optimal models. Rashomon sets can be extremely complicated, particularly for highly nonlinear function classes that allow complex interaction terms, such as decision trees. We provide the first technique for completely enumerating the Rashomon set for sparse decision trees; in fact, our work provides the first complete enumeration of any Rashomon set for a non-trivial problem with a highly nonlinear discrete function class. This allows the user an unprecedented level of control over model choice among all models that are approximately equally good. We represent the Rashomon set in a specialized data structure that supports efficient querying and sampling. We show three applications of the Rashomon set: 1) it can be used to study variable importance for the set of almost-optimal trees (as opposed to a single tree), 2) the Rashomon set for accuracy enables enumeration of the Rashomon sets for balanced accuracy and F1-score, and 3) the Rashomon set for a full dataset can be used to produce Rashomon sets constructed with only subsets of the data set. Thus, we are able to examine Rashomon sets across problems with a new lens, enabling users to choose models rather than be at the mercy of an algorithm that produces only a single model.

## [A Universal Error Measure for Input Predictions Applied to Online Graph Problems](#)

- Giulia Bernardini · Alexander Lindermayr · Alberto Marchetti-Spaccamela · Nicole Megow · Leen Stougie · Michelle Sweering
- abstract@[open-review](#): We introduce a novel measure for quantifying the error in input predictions. The error is based on a minimum-cost hyperedge cover in a suitably defined hypergraph and provides a general template which we apply to online graph problems. The measure captures errors due to absent predicted requests as well as unpredicted actual requests; hence, predicted and actual inputs can be of arbitrary size. We achieve refined performance guarantees for previously studied network design problems in the online-list model, such as Steiner tree and facility location. Further, we initiate the study of learning-augmented algorithms for online routing problems, such as the traveling salesperson problem and dial-a-ride problem, where (transportation) requests arrive over time (online-time model). We provide a general algorithmic framework and we give error-dependent performance bounds that improve upon known worst-case barriers, when given accurate predictions, at the cost of slightly increased worst-case bounds when given predictions of arbitrary quality.

## [Zeroth-Order Negative Curvature Finding: Escaping Saddle Points without Gradients](#)

- Hualin Zhang · Huan Xiong · Bin Gu
- abstract@[open-review](#): We consider escaping saddle points of nonconvex problems where only the function evaluations can be accessed. Although a variety of works have been proposed, the majority of them require either second or first-order information, and only a few of them have exploited zeroth-order methods, particularly the technique of negative curvature finding with zeroth-order methods which has been proven to be the most efficient method for escaping saddle points. To fill this gap, in this paper, we propose two zeroth-order negative curvature finding frameworks that can replace Hessian-vector product computations without increasing the iteration complexity. We apply the proposed frameworks to ZO-GD, ZO-SGD, ZO-SCSG, ZO-SPIDER and prove that these ZO algorithms can converge to  $(\epsilon, \delta)$ -approximate second-order stationary points with less query complexity compared with prior zeroth-order works for finding local minima.

## [Personalized Online Federated Multi-Kernel Learning](#)

- Pouya M. Ghari · Yanning Shen
- abstract@[open-review](#): Multi-kernel learning (MKL) exhibits well-documented performance in online non-linear function approximation. Federated learning enables a group of learners (called clients) to train an MKL model on the data distributed among clients to perform online non-linear function approximation. There are some challenges in online federated MKL that need to be addressed: i) Communication efficiency especially when a large number of kernels are considered ii) Heterogeneous data distribution among clients. The present paper develops an algorithmic framework to enable clients to communicate with the server to send their updates with affordable communication cost while clients employ a large dictionary of kernels. Utilizing random feature (RF) approximation, the present paper proposes scalable online federated MKL algorithm. We prove that using the proposed online federated MKL algorithm, each client enjoys sub-linear regret with respect to the RF approximation of its best kernel in hindsight, which indicates that the proposed algorithm can effectively deal with heterogeneity of the data distributed among clients. Experimental results on real datasets showcase the advantages of the proposed algorithm compared with other online federated kernel learning ones.

## [Online Learning and Pricing for Network Revenue Management with Reusable Resources](#)

- Huiwen Jia · Cong Shi · Siqian Shen
- abstract@[open-review](#): We consider a price-based network revenue management problem with multiple products and multiple reusable resources. Each randomly arriving customer requests a product (service) that needs to occupy a sequence of reusable resources (servers). We adopt an incomplete information setting where the firm does not know the price-demand function for each product and the goal is to dynamically set prices of all products to maximize the total expected revenue of serving customers. We propose novel batched bandit learning algorithms for finding near-optimal pricing policies, and show that they admit a near-optimal cumulative regret bound of  $\tilde{O}(J\sqrt{XT})$ , where  $J$ ,  $X$ , and  $T$  are the numbers of products, candidate prices, and service periods, respectively. As part of our regret analysis, we develop the first finite-time mixing time analysis of an open network queueing system (i.e., the Jackson Network), which could be of independent interest. Our numerical studies show very promising results of the proposed approaches.

## Delving into OOD Detection with Vision-Language Representations

- Yifei Ming · Ziyang Cai · Jiuxiang Gu · Yiyou Sun · Wei Li · Yixuan Li
- abstract@[open-review](#): Recognizing out-of-distribution (OOD) samples is critical for machine learning systems deployed in the open world. The vast majority of OOD detection methods are driven by a single modality (e.g., either vision or language), leaving the rich information in multi-modal representations untapped. Inspired by the recent success of vision-language pre-training, this paper enriches the landscape of OOD detection from a single-modal to a multi-modal regime. Particularly, we propose Maximum Concept Matching (MCM), a simple yet effective zero-shot OOD detection method based on aligning visual features with textual concepts. We contribute in-depth analysis and theoretical insights to understand the effectiveness of MCM. Extensive experiments demonstrate that our proposed MCM achieves superior performance on a wide variety of real-world tasks. MCM with vision-language features outperforms a common baseline with pure visual features on a hard OOD task with semantically similar classes by 56.60% (FPR95).

## Distributed Online Convex Optimization with Compressed Communication

- Zhipeng Tu · Xi Wang · Yiguang Hong · Lei Wang · Deming Yuan · Guodong Shi
- abstract@[open-review](#): We consider a distributed online convex optimization problem when streaming data are distributed among computing agents over a connected communication network. Since the data are high-dimensional or the network is large-scale, communication load can be a bottleneck for the efficiency of distributed algorithms. To tackle this bottleneck, we apply the state-of-art data compression scheme to the fundamental GD-based distributed online algorithms. Three algorithms with difference-compressed communication are proposed for full information feedback (DC-DOGD), one-point bandit feedback (DC-DOBD), and two-point bandit feedback (DC-DO2BD), respectively. We obtain regret bounds explicitly in terms of the time horizon, compression ratio, decision dimension, agent number, and network parameters. Our algorithms are proved to be no-regret and match the same regret bounds, w.r.t. the time horizon, with their uncompressed versions for both convex and strongly convex losses. Numerical experiments are given to validate the theoretical findings and illustrate that the proposed algorithms can effectively reduce the total transmitted bits for distributed online training compared with the uncompressed baseline.

## Benign, Tempered, or Catastrophic: Toward a Refined Taxonomy of Overfitting

- Neil Mallinar · James Simon · Amirhesam Abedsoltan · Parthe Pandit · Misha Belkin · Preetum Nakkiran
- abstract@[open-review](#): The practical success of overparameterized neural networks has motivated the recent scientific study of \emph{interpolating methods}— learning methods which are able fit their training data perfectly. Empirically, certain interpolating methods can fit noisy training data without catastrophically bad test performance, which defies standard intuitions from statistical learning theory. Aiming to explain this, a large body of recent work has studied \emph{benign overfitting}, a behavior seen in certain asymptotic settings under which interpolating methods approach Bayes-optimality, even in the presence of noise. In this work, we argue that, while benign overfitting has been instructive to study, real interpolating methods like deep networks do not fit benignly. That is, noise in the train set leads to suboptimal generalization, suggesting that these methods fall in an intermediate regime between benign and catastrophic overfitting, in which asymptotic risk is neither Bayes-optimal nor unbounded, with the confounding effect of the noise being ``tempered'' but non-negligible. We call this behavior \textit{tempered overfitting}. We first provide broad empirical evidence for our three-part taxonomy, demonstrating that deep neural networks and kernel machines fit to noisy data can be reasonably well classified as benign, tempered, or catastrophic. We then specialize to kernel (ridge) regression (KR), obtaining conditions on the ridge parameter and kernel eigenspectrum under which KR exhibits each of the three behaviors, demonstrating the consequences for KR with common kernels and trained neural networks of infinite width using experiments on natural and synthetic datasets.

## Molecule Generation by Principal Subgraph Mining and Assembling

- Xiangzhe Kong · Wenbing Huang · Zhixing Tan · Yang Liu
- abstract@[open-review](#): Molecule generation is central to a variety of applications. Current attention has been paid to approaching the generation task as subgraph prediction and assembling. Nevertheless, these methods usually rely on hand-crafted or external subgraph construction, and the subgraph assembling depends solely on local arrangement. In this paper, we define a novel notion, principal subgraph that is closely related to the informative pattern within molecules. Interestingly, our proposed merge-and-update subgraph extraction method can automatically discover frequent principal subgraphs from the dataset, while previous methods are incapable of. Moreover, we develop a two-step subgraph assembling strategy, which first predicts a set of subgraphs in a sequence-wise manner and then assembles all generated subgraphs globally as the final output molecule. Built upon graph variational auto-encoder, our model is demonstrated to be effective in terms of several evaluation metrics and efficiency, compared with state-of-the-art methods on distribution learning and (constrained) property optimization tasks.

## Size and depth of monotone neural networks: interpolation and approximation

- Dan Mikulincer · Daniel Reichman
- abstract@[open-review](#): Monotone functions and data sets arise in a variety of applications. We study the interpolation problem for monotone data sets: The input is a monotone data set with  $n$  points, and the goal is to find a size and depth efficient monotone neural network with \emph{non negative parameters} and threshold units that interpolates the data set. We show that there are monotone data sets that cannot be interpolated by a monotone network of depth 2. On the other hand, we prove that for every monotone data set with  $n$  points in  $\mathbb{R}^d$ , there exists an interpolating monotone network of depth 4 and size  $O(nd)$ . Our interpolation result implies that every monotone function over  $[0,1]^d$  can be approximated arbitrarily well by a depth-4 monotone network, improving the previous best-known construction of depth  $d+1$ . Finally, building on results from Boolean circuit complexity, we show that the inductive bias of having positive parameters can lead to a super-polynomial blow-up in the number of neurons when approximating monotone functions.

## Neural Approximation of Extended Persistent Homology on Graphs

- Zuoyu Yan · Tengfei Ma · Liangcai Gao · Zhi Tang · Yusu Wang · Chao Chen
- abstract@[open-review](#): Topological features based on persistent homology capture high-order structural information so as to augment graph neural network methods. However, computing extended persistent homology summaries remains slow for large and dense graphs and can be a serious bottleneck for the learning pipeline. Inspired by recent success in neural algorithmic reasoning, we propose a novel graph neural network to estimate extended persistence diagrams (EPDs) on graphs efficiently. Our model is built on algorithmic insights, and benefits from better supervision and closer alignment with the EPD computation algorithm. We validate our method with convincing empirical results on approximating EPDs and downstream graph representation learning tasks. Our method is also efficient; on large and dense graphs, we accelerate the computation by nearly 100 times.

## Decentralized Gossip-Based Stochastic Bilevel Optimization over Communication Networks

- Shuguang Yang · Xuezhou Zhang · Mengdi Wang
- abstract@[open-review](#): Bilevel optimization have gained growing interests, with numerous applications found in meta learning, minimax games, reinforcement learning, and nested composition optimization. This paper studies the problem of distributed bilevel optimization over a network where agents can only communicate with neighbors, including examples from multi-task, multi-agent learning and federated learning. In this paper, we propose a

gossip-based distributed bilevel learning algorithm that allows networked agents to solve both the inner and outer optimization problems in a single timescale and share information via network propagation. We show that our algorithm enjoys the  $\mathcal{O}(\frac{1}{K}\epsilon^2)$  per-agent sample complexity for general nonconvex bilevel optimization and  $\mathcal{O}(\frac{1}{K}\epsilon)$  for strongly convex objective, achieving a speedup that scales linearly with the network size. The sample complexities are optimal in both  $\epsilon$  and  $K$ . We test our algorithm on the examples of hyperparameter tuning and decentralized reinforcement learning. Simulated experiments confirmed that our algorithm achieves the state-of-the-art training efficiency and test accuracy.

## [Communication Efficient Distributed Learning for Kernelized Contextual Bandits](#)

- Chuanhao Li · Huazheng Wang · Mengdi Wang · Hongning Wang
- abstract@[open-review](#): We tackle the communication efficiency challenge of learning kernelized contextual bandits in a distributed setting. Despite the recent advances in communication-efficient distributed bandit learning, existing solutions are restricted to simple models like multi-armed bandits and linear bandits, which hamper their practical utility. In this paper, instead of assuming the existence of a linear reward mapping from the features to the expected rewards, we consider non-linear reward mappings, by letting agents collaboratively search in a reproducing kernel Hilbert space (RKHS). This introduces significant challenges in communication efficiency as distributed kernel learning requires the transfer of raw data, leading to a communication cost that grows linearly w.r.t. time horizon  $T$ . We address this issue by equipping all agents to communicate via a common Nyström embedding that gets updated adaptively as more data points are collected. We rigorously proved that our algorithm can attain sub-linear rate in both regret and communication cost.

## [Exploration-Guided Reward Shaping for Reinforcement Learning under Sparse Rewards](#)

- Rati Devidze · Parameswaran Kamalaruban · Adish Singla
- abstract@[open-review](#): We study the problem of reward shaping to accelerate the training process of a reinforcement learning agent. Existing works have considered a number of different reward shaping formulations; however, they either require external domain knowledge or fail in environments with extremely sparse rewards. In this paper, we propose a novel framework, Exploration-Guided Reward Shaping (ExploRS), that operates in a fully self-supervised manner and can accelerate an agent's learning even in sparse-reward environments. The key idea of ExploRS is to learn an intrinsic reward function in combination with exploration-based bonuses to maximize the agent's utility w.r.t. extrinsic rewards. We theoretically showcase the usefulness of our reward shaping framework in a special family of MDPs. Experimental results on several environments with sparse/noisy reward signals demonstrate the effectiveness of ExploRS.

## [DReS-FL: Dropout-Resilient Secure Federated Learning for Non-IID Clients via Secret Data Sharing](#)

- Jiawei Shao · Yuchang Sun · Songze Li · Jun Zhang
- abstract@[open-review](#): Federated learning (FL) strives to enable collaborative training of machine learning models without centrally collecting clients' private data. Different from centralized training, the local datasets across clients in FL are non-independent and identically distributed (non-IID). In addition, the data-owning clients may drop out of the training process arbitrarily. These characteristics will significantly degrade the training performance. This paper proposes a Dropout-Resilient Secure Federated Learning (DReS-FL) framework based on Lagrange coded computing (LCC) to tackle both the non-IID and dropout problems. The key idea is to utilize Lagrange coding to secretly share the private datasets among clients so that the effects of non-IID distribution and client dropouts can be compensated during local gradient computations. To provide a strict privacy guarantee for local datasets and correctly decode the gradient at the server, the gradient has to be a polynomial function in a finite field, and thus we construct polynomial integer neural networks (PINNs) to enable our framework. Theoretical analysis shows that DReS-FL is resilient to client dropouts and provides privacy protection for the local datasets. Furthermore, we experimentally demonstrate that DReS-FL consistently leads to significant performance gains over baseline methods.

## [Mildly Conservative Q-Learning for Offline Reinforcement Learning](#)

- Jiafei Lyu · Xiaoteng Ma · Xiu Li · Zongqing Lu
- abstract@[open-review](#): Offline reinforcement learning (RL) defines the task of learning from a static logged dataset without continually interacting with the environment. The distribution shift between the learned policy and the behavior policy makes it necessary for the value function to stay conservative such that out-of-distribution (OOD) actions will not be severely overestimated. However, existing approaches, penalizing the unseen actions or regularizing with the behavior policy, are too pessimistic, which suppresses the generalization of the value function and hinders the performance improvement. This paper explores mild but enough conservatism for offline learning while not harming generalization. We propose Mildly Conservative Q-learning (MCQ), where OOD actions are actively trained by assigning them proper pseudo Q values. We theoretically show that MCQ induces a policy that behaves at least as well as the behavior policy and no erroneous overestimation will occur for OOD actions. Experimental results on the D4RL benchmarks demonstrate that MCQ achieves remarkable performance compared with prior work. Furthermore, MCQ shows superior generalization ability when transferring from offline to online, and significantly outperforms baselines.

## [Data-Driven Model-Based Optimization via Invariant Representation Learning](#)

- Han Qi · Yi Su · Aviral Kumar · Sergey Levine
- abstract@[open-review](#): We study the problem of data-driven model-based optimization, where the goal is to find the optimal design, provided access to only a static dataset, with no active data collection. The central challenge in data-driven model-based optimization is distributional shift, where the optimizer is fooled into producing out-of-distribution (OOD) designs that erroneously appear promising under a model trained on the provided data. To address this issue, we formulate model-based optimization as domain adaptation, where the goal is to make accurate predictions for the value of designs during optimization ("target domain"), when training only on the dataset ("source domain"). This perspective leads to invariant objective models (IOM), our approach for addressing distributional shift by enforcing invariance between the learned representations of the training dataset and optimized designs. In IOM, if the optimized designs are too different from the training dataset, the representation will be forced to lose much of the information that distinguishes good designs from bad ones, making all choices seem mediocre. Critically, when the optimizer is aware of this representational tradeoff, it should choose not to stray too far from the training distribution, leading to a natural trade-off between distributional shift and learning performance.

## [Unpacking Reward Shaping: Understanding the Benefits of Reward Engineering on Sample Complexity](#)

- Abhishek Gupta · Aldo Pacchiano · Yuexiang Zhai · Sham Kakade · Sergey Levine
- abstract@[open-review](#): The success of reinforcement learning in a variety of challenging sequential decision-making problems has been much discussed, but often ignored in this discussion is the consideration of how the choice of reward function affects the behavior of these algorithms. Most practical RL algorithms require copious amounts of reward engineering in order to successfully solve challenging tasks. The idea of this type of ``reward-shaping'' has been often discussed in the literature and is used in practical instantiations, but there is relatively little formal characterization of how the choice of reward shaping can yield benefits in sample complexity for RL problems. In this work, we build on the framework of novelty-based exploration to provide a simple scheme for incorporating shaped rewards into RL along with an analysis tool to show that particular choices of reward shaping provably improve sample efficiency. We characterize the class of problems where these gains are expected to be significant and show how this can be connected to practical algorithms in the literature. We show that these results hold in practice in experimental evaluations as well, providing an insight into the mechanisms through which reward shaping can significantly improve the complexity of reinforcement learning while retaining asymptotic performance.

## SemMAE: Semantic-Guided Masking for Learning Masked Autoencoders

- Gang Li · Heliang Zheng · Daqing Liu · Chaoyue Wang · Bing Su · Changwen Zheng
- abstract@[open-review](#): Recently, significant progress has been made in masked image modeling to catch up to masked language modeling. However, unlike words in NLP, the lack of semantic decomposition of images still makes masked autoencoding (MAE) different between vision and language. In this paper, we explore a potential visual analogue of words, i.e., semantic parts, and we integrate semantic information into the training process of MAE by proposing a Semantic-Guided Masking strategy. Compared to widely adopted random masking, our masking strategy can gradually guide the network to learn various information, i.e., from intra-part patterns to inter-part relations. In particular, we achieve this in two steps. 1) Semantic part learning: we design a self-supervised part learning method to obtain semantic parts by leveraging and refining the multi-head attention of a ViT-based encoder. 2) Semantic-guided MAE (SemMAE) training: we design a masking strategy that varies from masking a portion of patches in each part to masking a portion of (whole) parts in an image. Extensive experiments on various vision tasks show that SemMAE can learn better image representation by integrating semantic information. In particular, SemMAE achieves 84.5% fine-tuning accuracy on ImageNet-1k, which outperforms the vanilla MAE by 1.4%. In the semantic segmentation and fine-grained recognition tasks, SemMAE also brings significant improvements and yields the state-of-the-art performance.

## Distinguishing discrete and continuous behavioral variability using warped autoregressive HMMs

- Julia Costacurta · Lea Duncker · Blue Sheffer · Alex Williams · Winthrop Gillis · Caleb Weinreb · Jeffrey Markowitz · Sandeep R Datta · Scott Linderman
- abstract@[open-review](#): A core goal in systems neuroscience and neuroethology is to understand how neural circuits generate naturalistic behavior. One foundational idea is that complex naturalistic behavior may be composed of sequences of stereotyped behavioral syllables, which combine to generate rich sequences of actions. To investigate this, a common approach is to use autoregressive hidden Markov models (ARHMMs) to segment video into discrete behavioral syllables. While these approaches have been successful in extracting syllables that are interpretable, they fail to account for other forms of behavioral variability, such as differences in speed, which may be better described as continuous in nature. To overcome these limitations, we introduce a class of warped ARHMMs (WARHMM). As is the case in the ARHMM, behavior is modeled as a mixture of autoregressive dynamics. However, the dynamics under each discrete latent state (i.e. each behavioral syllable) are additionally modulated by a continuous latent ``warping variable.'' We present two versions of warped ARHMM in which the warping variable affects the dynamics of each syllable either linearly or nonlinearly. Using depth-camera recordings of freely moving mice, we demonstrate that the failure of ARHMMs to account for continuous behavioral variability results in duplicate cluster assignments. WARHMM achieves similar performance to the standard ARHMM while using fewer behavioral syllables. Further analysis of behavioral measurements in mice demonstrates that WARHMM identifies structure relating to response vigor.

## Learning to Drop Out: An Adversarial Approach to Training Sequence VAEs

- Djordje Miladinovic · Kumar Shridhar · Kushal Jain · Max Paulus · Joachim M Buhmann · Carl Allen
- abstract@[open-review](#): In principle, applying variational autoencoders (VAEs) to sequential data offers a method for controlled sequence generation, manipulation, and structured representation learning. However, training sequence VAEs is challenging: autoregressive decoders can often explain the data without utilizing the latent space, known as posterior collapse. To mitigate this, state-of-the-art models weaken the powerful decoder by applying uniformly random dropout to the decoder input. We show theoretically that this removes pointwise mutual information provided by the decoder input, which is compensated for by utilizing the latent space. We then propose an adversarial training strategy to achieve information-based stochastic dropout. Compared to uniform dropout on standard text benchmark datasets, our targeted approach increases both sequence modeling performance and the information captured in the latent space.

## Diffusion-based Molecule Generation with Informative Prior Bridges

- Lemeng Wu · Chengyue Gong · Xingchao Liu · Mao Ye · Qiang Liu
- abstract@[open-review](#): AI-based molecule generation provides a promising approach to a large area of biomedical sciences and engineering, such as antibody design, hydrolase engineering, or vaccine development. Because the molecules are governed by physical laws, a key challenge is to incorporate prior information into the training procedure to generate high-quality and realistic molecules. We propose a simple and novel approach to steer the training of diffusion-based generative models with physical and statistics prior information. This is achieved by constructing physically informed diffusion bridges, stochastic processes that guarantee to yield a given observation at the fixed terminal time. We develop a Lyapunov function based method to construct and determine bridges, and propose a number of proposals of informative prior bridges for both high-quality molecule generation and uniformity-promoted 3D point cloud generation. With comprehensive experiments, we show that our method provides a powerful approach to the 3D generation task, yielding molecule structures with better quality and stability scores and more uniformly distributed point clouds of high qualities.

## Understanding the Generalization Benefit of Normalization Layers: Sharpness Reduction

- Kaifeng Lyu · Zhiyuan Li · Sanjeev Arora
- abstract@[open-review](#): Normalization layers (e.g., Batch Normalization, Layer Normalization) were introduced to help with optimization difficulties in very deep nets, but they clearly also help generalization, even in not-so-deep nets. Motivated by the long-held belief that flatter minima lead to better generalization, this paper gives mathematical analysis and supporting experiments suggesting that normalization (together with accompanying weight-decay) encourages GD to reduce the sharpness of loss surface. Here ``sharpness'' is carefully defined given that the loss is scale-invariant, a known consequence of normalization. Specifically, for a fairly broad class of neural nets with normalization, our theory explains how GD with a finite learning rate enters the so-called Edge of Stability (EoS) regime, and characterizes the trajectory of GD in this regime via a continuous sharpness-reduction flow.

## Coresets for Relational Data and The Applications

- Jiaxiang Chen · Qingyuan Yang · Ruomin Huang · Hu Ding
- abstract@[open-review](#): A coreset is a small set that can approximately preserve the structure of the original input data set. Therefore we can run our algorithm on a coreset so as to reduce the total computational complexity. Conventional coresets techniques assume that the input data set is available to process explicitly. However, this assumption may not hold in real-world scenarios. In this paper, we consider the problem of coressets construction over relational data. Namely, the data is decoupled into several relational tables, and it is expensive to directly materialize the data matrix by joining the tables. We propose a novel approach called ``aggregation tree with pseudo-cube'' that can build a coreset from bottom to top. Moreover, our approach can neatly circumvent several troublesome issues of relational learning problems [Khamis et al., PODS 2019]. Under some mild assumptions, we show that our coreset approach can be applied for the machine learning tasks, such as clustering, logistic regression and SVM.

## Shield Decentralization for Safe Multi-Agent Reinforcement Learning

- Daniel Melcer · Stavros Tripakis · Christopher Amato
- abstract@[open-review](#): Learning safe solutions is an important but challenging problem in multi-agent reinforcement learning (MARL). Shielded reinforcement learning is one approach for preventing agents from choosing unsafe actions. Current shielded reinforcement learning methods for MARL make strong assumptions about communication and full observability. In this work, we extend the formalization of the shielded reinforcement learning problem to a truly decentralized multi-agent setting. We then present an algorithm for decomposition of a centralized shield, allowing shields to be used in

such decentralized, communication-free environments. Our results show that agents equipped with decentralized shields perform comparably to agents with centralized shields in several tasks, allowing shielding to be used in decentralized environments for the first time.

## [Forecasting Human Trajectory from Scene History](#)

- Mancheng Meng · Ziyan Wu · Terrence Chen · Dinggang Shen · Fan Yang
- abstract@[open-review](#): Predicting the future trajectory of a person remains a challenging problem, due to randomness and subjectivity. However, the moving patterns of human in constrained scenario typically conform to a limited number of regularities to a certain extent, because of the scenario restrictions (e.g., floor plan, roads and obstacles) and person-person or person-object interactivity. Thus, an individual person in this scenario should follow one of the regularities as well. In other words, a person's subsequent trajectory has likely been traveled by others. Based on this hypothesis, we propose to forecast a person's future trajectory by learning from the implicit scene regularities. We call the regularities, inherently derived from the past dynamics of the people and the environment in the scene, *scene history*. We categorize scene history information into two types: historical group trajectories and individual-surroundings interaction. To exploit these information for trajectory prediction, we propose a novel framework Scene History Excavating Network (SHENet), where the scene history is leveraged in a simple yet effective approach. In particular, we design two components, the group trajectory bank module to extract representative group trajectories as the candidate for future path, and the cross-modal interaction module to model the interaction between individual past trajectory and its surroundings for trajectory refinement, respectively. In addition, to mitigate the uncertainty in the evaluation, caused by the aforementioned randomness and subjectivity, we propose to include smoothness into evaluation metrics. We conduct extensive evaluations to validate the efficacy of proposed framework on ETH, UCY, as well as a new, challenging benchmark dataset PAV, demonstrating superior performance compared to state-of-the-art methods.

## [A general approximation lower bound in \$L^p\$ norm, with applications to feedforward neural networks](#)

- El Mehdi Achour · Armand Foucault · Sébastien Gerchinovitz · François Malgouyres
- abstract@[open-review](#): We study the fundamental limits to the expressive power of neural networks. Given two sets  $\mathcal{F}$ ,  $\mathcal{G}$  of real-valued functions, we first prove a general lower bound on how well functions in  $\mathcal{F}$  can be approximated in  $L^p(\mu)$  norm by functions in  $\mathcal{G}$ , for any  $p \geq 1$  and any probability measure  $\mu$ . The lower bound depends on the packing number of  $\mathcal{F}$ , the range of  $\mathcal{F}$ , and the fat-shattering dimension of  $\mathcal{G}$ . We then instantiate this bound to the case where  $\mathcal{G}$  corresponds to a piecewise-polynomial feedforward neural network, and describe in details the application to two sets  $\mathcal{F}$ : Hölder balls and multivariate monotonic functions. Beside matching (known or new) upper bounds up to log factors, our lower bounds shed some light on the similarities or differences between approximation in  $L^p$  norm or in sup norm, solving an open question by DeVore et al. (2021). Our proof strategy differs from the sup norm case and uses a key probability result of Mendelson (2002).

## [Do Current Multi-Task Optimization Methods in Deep Learning Even Help?](#)

- Derrick Xin · Behrooz Ghorbani · Justin Gilmer · Ankush Garg · Orhan Firat
- abstract@[open-review](#): Recent research has proposed a series of specialized optimization algorithms for deep multi-task models. It is often claimed that these multi-task optimization (MTO) methods yield solutions that are superior to the ones found by simply optimizing a weighted average of the task losses. In this paper, we perform large-scale experiments on a variety of language and vision tasks to examine the empirical validity of these claims. We show that, despite the added design and computational complexity of these algorithms, MTO methods do not yield any performance improvements beyond what is achievable via traditional optimization approaches. We highlight alternative strategies that consistently yield improvements to the performance profile and point out common training pitfalls that might cause suboptimal results. Finally, we outline challenges in reliably evaluating the performance of MTO algorithms and discuss potential solutions.

## [ConfounderGAN: Protecting Image Data Privacy with Causal Confounder](#)

- Qi Tian · Kelu Jiang · Kun Kuang · Furui Liu · Zhihua Wang · Fei Wu
- abstract@[open-review](#): The success of deep learning is partly attributed to the availability of massive data downloaded freely from the Internet. However, it also means that users' private data may be collected by commercial organizations without consent and used to train their models. Therefore, it's important and necessary to develop a method or tool to prevent unauthorized data exploitation. In this paper, we propose ConfounderGAN, a generative adversarial network (GAN) that can make personal image data unlearnable to protect the data privacy of its owners. Specifically, the noise produced by the generator for each image has the confounder property. It can build spurious correlations between images and labels, so that the model cannot learn the correct mapping from images to labels in this noise-added dataset. Meanwhile, the discriminator is used to ensure that the generated noise is small and imperceptible, thereby remaining the normal utility of the encrypted image for humans. The experiments are conducted in six image classification datasets, including three natural object datasets and three medical datasets. The results demonstrate that our method not only outperforms state-of-the-art methods in standard settings, but can also be applied to fast encryption scenarios. Moreover, we show a series of transferability and stability experiments to further illustrate the effectiveness and superiority of our method.

## [A Fair Comparison of Two Popular Flat-Minima Optimizers: Stochastic Weight Averaging vs. Sharpness-Aware Minimization](#)

- Jean Kaddour · Linqing Liu · Ricardo Silva · Matt Kusner
- abstract@[open-review](#): Recently, flat-minima optimizers, which seek to find parameters in low loss neighborhoods, have been shown to improve upon stochastic and adaptive gradient-based optimizers for training neural networks. Two methods have received significant attention due to their impressive generalization performance and scalability: 1. Stochastic Weight Averaging (SWA), and 2. Sharpness Aware Minimization (SAM). However, despite this, there has been limited investigation into their different properties and no systematic benchmarking of them. Previous work mainly evaluated SWA and SAM on different architectures and datasets. We fill this gap here by comparing the loss surfaces of the models trained with each method and through a broad benchmarking across computer vision, natural language processing, and graph representation learning tasks. We discover a number of surprising findings from these results, which we hope will help researchers further improve deep learning optimizers, and practitioners identify the right optimizer for their problem.

## [On the Parameterization and Initialization of Diagonal State Space Models](#)

- Albert Gu · Karan Goel · Ankit Gupta · Christopher Rafferty
- abstract@[open-review](#): State space models (SSM) have recently been shown to be very effective as a deep learning layer as a promising alternative to sequence models such as RNNs, CNNs, or Transformers. The first version to show this potential was the S4 model, which is particularly effective on tasks involving long-range dependencies by using a prescribed state matrix called the HiPPO matrix. While this has an interpretable mathematical mechanism for modeling long dependencies, it also requires a custom representation and algorithm that makes the model difficult to understand and implement. On the other hand, a recent variant of S4 called DSS showed that restricting the state matrix to be fully diagonal can still preserve the performance of the original model when using a specific initialization based on approximating S4's matrix. This work seeks to systematically understand how to parameterize and initialize diagonal state space models. While it follows from classical results that almost all SSMs have an equivalent diagonal form, we show that the initialization is critical for performance. First, we explain why DSS works mathematically, as the diagonal approximation to S4 surprisingly recovers the same dynamics in the limit of infinite state dimension. We then systematically describe various design choices in parameterizing and computing diagonal SSMs, and perform a controlled empirical study ablating the effects of these choices. Our final model S4D is a simple diagonal version of S4 whose kernel

computation requires just 3 lines of code and performs comparably to S4 in almost all settings, with state-of-the-art results in image, audio, and medical time-series domains, and 85% average on the Long Range Arena benchmark.

## [Single-Stage Visual Relationship Learning using Conditional Queries](#)

- Alakh Desai · Tz-Ying Wu · Subarna Tripathi · Nuno Vasconcelos
- abstract@[open-review](#): Research in scene graph generation (SGG) usually considers two-stage models, that is, detecting a set of entities, followed by combining them and labeling all possible relationships. While showing promising results, the pipeline structure induces large parameter and computation overhead, and typically hinders end-to-end optimizations. To address this, recent research attempts to train single-stage models that are more computationally efficient. With the advent of DETR, a set-based detection model, one-stage models attempt to predict a set of subject-predicate-object triplets directly in a single shot. However, SGG is inherently a multi-task learning problem that requires modeling entity and predicate distributions simultaneously. In this paper, we propose Transformers with conditional queries for SGG, namely, TraCQ with a new formulation for SGG that avoids the multi-task learning problem and the combinatorial entity pair distribution. We employ a DETR-based encoder-decoder design and leverage conditional queries to significantly reduce the entity label space as well, which leads to 20% fewer parameters compared to state-of-the-art one-stage models. Experimental results show that TraCQ not only outperforms existing single-stage scene graph generation methods, it also beats state-of-the-art two-stage methods on the Visual Genome dataset, yet is capable of end-to-end training and faster inference.

## [Optimizing Data Collection for Machine Learning](#)

- Rafid Mahmood · James Lucas · Jose M. Alvarez · Sanja Fidler · Marc Law
- abstract@[open-review](#): Modern deep learning systems require huge data sets to achieve impressive performance, but there is little guidance on how much or what kind of data to collect. Over-collecting data incurs unnecessary present costs, while under-collecting may incur future costs and delay workflows. We propose a new paradigm for modeling the data collection workflow as a formal optimal data collection problem that allows designers to specify performance targets, collection costs, a time horizon, and penalties for failing to meet the targets. Additionally, this formulation generalizes to tasks requiring multiple data sources, such as labeled and unlabeled data used in semi-supervised learning. To solve our problem, we develop Learn-Optimize-Collect (LOC), which minimizes expected future collection costs. Finally, we numerically compare our framework to the conventional baseline of estimating data requirements by extrapolating from neural scaling laws. We significantly reduce the risks of failing to meet desired performance targets on several classification, segmentation, and detection tasks, while maintaining low total collection costs.

## [Convergent Representations of Computer Programs in Human and Artificial Neural Networks](#)

- Shashank Srikant · Ben Lipkin · Anna Ivanova · Evelina Fedorenko · Una-May O'Reilly
- abstract@[open-review](#): What aspects of computer programs are represented by the human brain during comprehension? We leverage brain recordings derived from functional magnetic resonance imaging (fMRI) studies of programmers comprehending Python code to evaluate the properties and code-related information encoded in the neural signal. We first evaluate a selection of static and dynamic code properties, such as abstract syntax tree (AST)-related and runtime-related metrics. Then, to learn whether brain representations encode fine-grained information about computer programs, we train a probe to align brain recordings with representations learned by a suite of ML models. We find that both the Multiple Demand and Language systems--brain systems which are responsible for very different cognitive tasks, encode specific code properties and uniquely align with machine learned representations of code. These findings suggest at least two distinct neural mechanisms mediating computer program comprehension and evaluation, prompting the design of code model objectives that go beyond static language modeling. We make all the corresponding code, data, and analysis publicly available.

## [Conformalized Fairness via Quantile Regression](#)

- Meichen Liu · Lei Ding · Dengdeng Yu · Wulong Liu · Linglong Kong · Bei Jiang
- abstract@[open-review](#): Algorithmic fairness has received increased attention in socially sensitive domains. While rich literature on mean fairness has been established, research on quantile fairness remains sparse but vital. To fulfill great needs and advocate the significance of quantile fairness, we propose a novel framework to learn a real-valued quantile function under the fairness requirement of Demographic Parity with respect to sensitive attributes, such as race or gender, and thereby derive a reliable fair prediction interval. Using optimal transport and functional synchronization techniques, we establish theoretical guarantees of distribution-free coverage and exact fairness for the induced prediction interval constructed by fair quantiles. A hands-on pipeline is provided to incorporate flexible quantile regressions with an efficient fairness adjustment post-processing algorithm. We demonstrate the superior empirical performance of this approach on several benchmark datasets. Our results show the model's ability to uncover the mechanism underlying the fairness-accuracy trade-off in a wide range of societal and medical applications.

## [A General Framework for Auditing Differentially Private Machine Learning](#)

- Fred Lu · Joseph Munoz · Maya Fuchs · Tyler LeBlond · Elliott Zaresky-Williams · Edward Raff · Francis Ferraro · Brian Testa
- abstract@[open-review](#): We present a framework to statistically audit the privacy guarantee conferred by a differentially private machine learner in practice. While previous works have taken steps toward evaluating privacy loss through poisoning attacks or membership inference, they have been tailored to specific models or have demonstrated low statistical power. Our work develops a general methodology to empirically evaluate the privacy of differentially private machine learning implementations, combining improved privacy search and verification methods with a toolkit of influence-based poisoning attacks. We demonstrate significantly improved auditing power over previous approaches on a variety of models including logistic regression, Naive Bayes, and random forest. Our method can be used to detect privacy violations due to implementation errors or misuse. When violations are not present, it can aid in understanding the amount of information that can be leaked from a given dataset, algorithm, and privacy specification.

## [Refined Dimension-Dependent Analysis for Private Convex Learning and Implications for Fine-Tuning](#)

- Xuechen Li · Daogao Liu · Tatsunori Hashimoto · Huseyin A. Inan · Janardhan Kulkarni · Yin-Tat Lee · Abhradeep Guha Thakurta
- abstract@[open-review](#): Large pretrained models can be privately fine-tuned to achieve performance approaching non-private models. A common theme in these results is the surprising observation that high-dimensional models can achieve favorable privacy-utility trade-offs. This seemingly contradicts known results on the model-size dependence of differentially private convex learning and raises the following research question: When does the performance of differentially private learning not degrade with increasing model size? We identify that the magnitudes of gradients projected onto subspaces is a key factor that determines performance. To precisely characterize this for private convex learning, we introduce a condition on the objective that we term \texttt{restricted Lipschitz continuity} and derive refined bounds for the excess empirical and population risks that are dimension-independent under additional conditions. We empirically show that in private fine-tuning of large language models, gradients obtained during fine-tuning are mostly controlled by a few principal components. This behavior is similar to conditions under which we obtain dimension-independent bounds in convex settings and provides a possible explanation for recent successes in large-scale private fine-tuning.

## [Autoformalization with Large Language Models](#)

- Yuhuai Wu · Albert Qiaoju Jiang · Wenda Li · Markus N Rabe · Charles Staats · Mateja Jamnik · Christian Szegedy

- abstract@[open-review](#): Autoformalization is the process of automatically translating from natural language mathematics to formal specifications and proofs. A successful autoformalization system could advance the fields of formal verification, program synthesis, and artificial intelligence. While the long-term goal of autoformalization seemed elusive for a long time, we show large language models provide new prospects towards this goal. We make the surprising observation that LLMs can correctly translate a significant portion (25.3%) of mathematical competition problems perfectly to formal specifications in Isabelle/HOL. We demonstrate the usefulness of this process by improving a previously introduced neural theorem prover via training on these autoformalized theorems. Our methodology results in a new state-of-the-art result on the MiniF2F theorem proving benchmark, improving the proof rate from \$29.6\%\$ to \$35.2\%\$.

## [A contrastive rule for meta-learning](#)

- Nicolas Zucchetti · Simon Schug · Johannes von Oswald · Dominic Zhao · João Sacramento
- abstract@[open-review](#): Humans and other animals are capable of improving their learning performance as they solve related tasks from a given problem domain, to the point of being able to learn from extremely limited data. While synaptic plasticity is generically thought to underlie learning in the brain, the precise neural and synaptic mechanisms by which learning processes improve through experience are not well understood. Here, we present a general-purpose, biologically-plausible meta-learning rule which estimates gradients with respect to the parameters of an underlying learning algorithm by simply running it twice. Notably, our rule neither requires computing second derivatives nor going backwards in time, two characteristic features of previous gradient-based methods that are hard to conceive in physical neural circuits. We demonstrate the generality of our rule by applying it to two distinct models: a complex synapse with internal states which consolidate task-shared information, and a dual-system architecture in which a primary network is rapidly modulated by another one to learn the specifics of each task. For both models, our meta-learning rule matches or outperforms reference algorithms on a wide range of benchmark problems, while only using information presumed to be locally available at neurons and synapses. We corroborate these findings with a theoretical analysis of the gradient estimation error incurred by our rule.

## [Quo Vadis: Is Trajectory Forecasting the Key Towards Long-Term Multi-Object Tracking?](#)

- Patrick Dendorfer · Vladimir Yugay · Aljosa Osep · Laura Leal-Taix©
- abstract@[open-review](#): Recent developments in monocular multi-object tracking have been very successful in tracking visible objects and bridging short occlusion gaps, mainly relying on data-driven appearance models. While we have significantly advanced short-term tracking performance, bridging longer occlusion gaps remains elusive: state-of-the-art object trackers bridge less than 10% of occlusions longer than three seconds. We suggest that the missing key is reasoning about future trajectories over a longer time horizon. Intuitively, the longer the occlusion gap, the larger the search space for possible associations. In this paper, we show that even a small yet diverse set of trajectory predictions for moving agents will significantly reduce this search space and thus improve long-term tracking robustness. Furthermore, we show that crucial components of our approach are reasoning in bird-eye space and generating a small yet diverse set of forecasts while accounting for their localization uncertainty. This way, we can advance state-of-the-art trackers on the MOTChallenge dataset and show we can significantly improve their long-term tracking performance.

## [Posterior Refinement Improves Sample Efficiency in Bayesian Neural Networks](#)

- Agustinus Kristiadi · Runa Eschenhagen · Philipp Hennig
- abstract@[open-review](#): Monte Carlo (MC) integration is the de facto method for approximating the predictive distribution of Bayesian neural networks (BNNs). But, even with many MC samples, Gaussian-based BNNs could still yield bad predictive performance due to the posterior approximation's error. Meanwhile, alternatives to MC integration are expensive. In this work, we experimentally show that the key to good MC-approximated predictive distributions is the quality of the approximate posterior itself. However, previous methods for obtaining accurate posterior approximations are expensive and non-trivial to implement. We, therefore, propose to refine Gaussian approximate posteriors with normalizing flows. When applied to last-layer BNNs, it yields a simple, cost-efficient, post hoc method for improving pre-existing parametric approximations. We show that the resulting posterior approximation is competitive with even the gold-standard full-batch Hamiltonian Monte Carlo.

## [Invariance Learning in Deep Neural Networks with Differentiable Laplace Approximations](#)

- Alexander Immer · Tycho van der Ouderaa · Gunnar Rätsch · Vincent Fortuin · Mark van der Wilk
- abstract@[open-review](#): Data augmentation is commonly applied to improve performance of deep learning by enforcing the knowledge that certain transformations on the input preserve the output. Currently, the used data augmentation is chosen by human effort and costly cross-validation, which makes it cumbersome to apply to new datasets. We develop a convenient gradient-based method for selecting the data augmentation without validation data and during training of a deep neural network. Our approach relies on phrasing data augmentation as an invariance in the prior distribution and learning it using Bayesian model selection, which has been shown to work in Gaussian processes, but not yet for deep neural networks. We propose a differentiable Kronecker-factored Laplace approximation to the marginal likelihood as our objective, which can be optimised without human supervision or validation data. We show that our method can successfully recover invariances present in the data, and that this improves generalisation and data efficiency on image datasets.

## [Enhanced Meta Reinforcement Learning via Demonstrations in Sparse Reward Environments](#)

- Desik Rengarajan · Sapana Chaudhary · Jaewon Kim · Dileep Kalathil · Srinivas Shakkottai
- abstract@[open-review](#): Meta reinforcement learning (Meta-RL) is an approach wherein the experience gained from solving a variety of tasks is distilled into a meta-policy. The meta-policy, when adapted over only a small (or just a single) number of steps, is able to perform near-optimally on a new, related task. However, a major challenge to adopting this approach to solve real-world problems is that they are often associated with sparse reward functions that only indicate whether a task is completed partially or fully. We consider the situation where some data, possibly generated by a sub-optimal agent, is available for each task. We then develop a class of algorithms entitled Enhanced Meta-RL via Demonstrations (EMRLD) that exploit this information---even if sub-optimal---to obtain guidance during training. We show how EMRLD jointly utilizes RL and supervised learning over the offline data to generate a meta-policy that demonstrates monotone performance improvements. We also develop a warm started variant called EMRLD-WS that is particularly efficient for sub-optimal demonstration data. Finally, we show that our EMRLD algorithms significantly outperform existing approaches in a variety of sparse reward environments, including that of a mobile robot.

## [DARE: Disentanglement-Augmented Rationale Extraction](#)

- Linan Yue · Qi Liu · Yichao Du · Yanqing An · Li Wang · Enhong Chen
- abstract@[open-review](#): Rationale extraction can be considered as a straightforward method of improving the model explainability, where rationales are a subsequence of the original inputs, and can be extracted to support the prediction results. Existing methods are mainly cascaded with the selector which extracts the rationale tokens, and the predictor which makes the prediction based on selected tokens. Since previous works fail to fully exploit the original input, where the information of non-selected tokens is ignored, in this paper, we propose a Disentanglement-Augmented Rationale Extraction (DARE) method, which encapsulates more information from the input to extract rationales. Specifically, it first disentangles the input into the rationale representations and the non-rationale ones, and then learns more comprehensive rationale representations for extracting by minimizing the mutual information (MI) between the two disentangled representations. Besides, to improve the performance of MI minimization, we develop a new MI estimator by exploring existing MI estimation methods. Extensive experimental results on two real-world datasets and simulation studies clearly validate the effectiveness of our proposed method.

## [DetCLIP: Dictionary-Enriched Visual-Concept Paralleled Pre-training for Open-world Detection](#)

- Lewei Yao · Jianhua Han · Youpeng Wen · Xiaodan Liang · Dan Xu · Wei Zhang · Zhenguo Li · Chunjing XU · Hang Xu
- abstract@[open-review](#): Open-world object detection, as a more general and challenging goal, aims to recognize and localize objects described by arbitrary category names. The recent work GLIP formulates this problem as a grounding problem by concatenating all category names of detection datasets into sentences, which leads to inefficient interaction between category names. This paper presents DetCLIP, a paralleled visual-concept pre-training method for open-world detection by resorting to knowledge enrichment from a designed concept dictionary. To achieve better learning efficiency, we propose a novel paralleled concept formulation that extracts concepts separately to better utilize heterogeneous datasets (i.e., detection, grounding, and image-text pairs) for training. We further design a concept dictionary (with descriptions) from various online sources and detection datasets to provide prior knowledge for each concept. By enriching the concepts with their descriptions, we explicitly build the relationships among various concepts to facilitate the open-domain learning. The proposed concept dictionary is further used to provide sufficient negative concepts for the construction of the word-region alignment loss, and to complete labels for objects with missing descriptions in captions of image-text pair data. The proposed framework demonstrates strong zero-shot detection performances, e.g., on the LVIS dataset, our DetCLIP-T outperforms GLIP-T by 9.9% mAP and obtains a 13.5% improvement on rare categories compared to the fully-supervised model with the same backbone as ours.

## [Eliciting Thinking Hierarchy without a Prior](#)

- Yuqing Kong · Yunqi Li · Yubo Zhang · Zhihuan Huang · Jinzhao Wu
- abstract@[open-review](#): When we use the wisdom of the crowds, we usually rank the answers according to their popularity, especially when we cannot verify the answers. However, this can be very dangerous when the majority make systematic mistakes. A fundamental question arises: can we build a hierarchy among the answers without any prior where the higher-ranking answers, which may not be supported by the majority, are from more sophisticated people? To address the question, we propose 1) a novel model to describe people's thinking hierarchy; 2) two algorithms to learn the thinking hierarchy without any prior; 3) a novel open-response based crowdsourcing approach based on the above theoretic framework. In addition to theoretic justifications, we conduct four empirical crowdsourcing studies and show that a) the accuracy of the top-ranking answers learned by our approach is much higher than that of plurality voting (In one question, the plurality answer is supported by 74 respondents but the correct answer is only supported by 3 respondents. Our approach ranks the correct answer the highest without any prior); b) our model has a high goodness-of-fit, especially for the questions where our top-ranking answer is correct. To the best of our knowledge, we are the first to propose a thinking hierarchy model with empirical validations in the general problem-solving scenarios; and the first to propose a practical open-response-based crowdsourcing approach that beats plurality voting without any prior.

## [Multi-Class \\$H\\$-Consistency Bounds](#)

- Pranjal Awasthi · Anqi Mao · Mehryar Mohri · Yutao Zhong
- abstract@[open-review](#): We present an extensive study of \$H\$-consistency bounds for multi-class classification. These are upper bounds on the target loss estimation error of a predictor in a hypothesis set \$H\$, expressed in terms of the surrogate loss estimation error of that predictor. They are stronger and more significant guarantees than Bayes-consistency, \$H\$-calibration or \$H\$-consistency, and more informative than excess error bounds derived for \$H\$ being the family of all measurable functions. We give a series of new \$H\$-consistency bounds for surrogate multi-class losses, including max losses, sum losses, and constrained losses, both in the non-adversarial and adversarial cases, and for differentiable or convex auxiliary functions used. We also prove that a non-trivial \$H\$-consistency bound can be given in some cases. To our knowledge, these are the first \$H\$-consistency bounds proven for the multi-class setting. Our proof techniques are also novel and likely to be useful in the analysis of other such guarantees.

## [Human-Robotic Prosthesis as Collaborating Agents for Symmetrical Walking](#)

- Ruofan Wu · Junmin Zhong · Brent Wallace · Xiang Gao · He Huang · Jennie Si
- abstract@[open-review](#): This is the first attempt at considering human influence in the reinforcement learning control of a robotic lower limb prosthesis toward symmetrical walking in real world situations. We propose a collaborative multi-agent reinforcement learning (cMARL) solution framework for this highly complex and challenging human-prosthesis collaboration (HPC) problem. The design of an automatic controller of the robot within the HPC context is based on accessible physical features or measurements that are known to affect walking performance. Comparisons are made with the current state-of-the-art robot control designs, which are single-agent based, as well as existing MARL solution approaches tailored to the problem, including multi-agent deep deterministic policy gradient (MADDPG) and counterfactual multi-agent policy gradient (COMA). Results show that, when compared to these approaches, treating the human and robot as coupled agents and using estimated human adaption in robot control design can achieve lower stage cost, peak error, and symmetry value to ensure better human walking performance. Additionally, our approach accelerates learning of walking tasks and increases learning success rate. The proposed framework can potentially be further developed to examine how human and robotic lower limb prosthesis interact, an area that little is known about. Advancing cMARL toward real world applications such as HPC for normative walking sets a good example of how AI can positively impact on people's lives.

## [Transformer-based Working Memory for Multiagent Reinforcement Learning with Action Parsing](#)

- Yaodong Yang · Guangyong Chen · Weixun Wang · Xiaotian Hao · Jianye Hao · Pheng-Ann Heng
- abstract@[open-review](#): Learning in real-world multiagent tasks is challenging due to the usual partial observability of each agent. Previous efforts alleviate the partial observability by historical hidden states with Recurrent Neural Networks, however, they do not consider the multiagent characters that either the multiagent observation consists of a number of object entities or the action space shows clear entity interactions. To tackle these issues, we propose the Agent Transformer Memory (ATM) network with a transformer-based memory. First, ATM utilizes the transformer to enable the unified processing of the factored environmental entities and memory. Inspired by the human's working memory process where a limited capacity of information temporarily held in mind can effectively guide the decision-making, ATM updates its fixed-capacity memory with the working memory updating schema. Second, as agents' each action has its particular interaction entities in the environment, ATM parses the action space to introduce this action's semantic inductive bias by binding each action with its specified involving entity to predict the state-action value or logit. Extensive experiments on the challenging SMAC and Level-Based Foraging environments validate that ATM could boost existing multiagent RL algorithms with impressive learning acceleration and performance improvement.

## [Chaotic Dynamics are Intrinsic to Neural Network Training with SGD](#)

- Luis Herrmann · Maximilian Granz · Tim Landgraf
- abstract@[open-review](#): With the advent of deep learning over the last decade, a considerable amount of effort has gone into better understanding and enhancing Stochastic Gradient Descent so as to improve the performance and stability of artificial neural network training. Active research fields in this area include exploiting second order information of the loss landscape and improving the understanding of chaotic dynamics in optimization. This paper exploits the theoretical connection between the curvature of the loss landscape and chaotic dynamics in neural network training to propose a modified SGD ensuring non-chaotic training dynamics to study the importance thereof in NN training. Building on this, we present empirical evidence suggesting that the negative eigenspectrum - and thus directions of local chaos - cannot be removed from SGD without hurting training performance. Extending our empirical analysis to long-term chaos dynamics, we challenge the widespread understanding of convergence against a confined region in parameter space. Our results show that although chaotic network behavior is mostly confined to the initial training phase, models perturbed upon initialization do diverge at a slow pace even after reaching top training performance, and that their divergence can be modelled through a composition of a random walk and a linear

divergence. The tools and insights developed as part of our work contribute to improving the understanding of neural network training dynamics and provide a basis for future improvements of optimization methods.

## [On Convergence of FedProx: Local Dissimilarity Invariant Bounds, Non-smoothness and Beyond](#)

- Xiaotong Yuan · Ping Li
- abstract@[open-review](#): The \FedProx~algorithm is a simple yet powerful distributed proximal point optimization method widely used for federated learning (FL) over heterogeneous data. Despite its popularity and remarkable success witnessed in practice, the theoretical understanding of FedProx is largely underinvestigated: the appealing convergence behavior of \FedProx~is so far characterized under certain non-standard and unrealistic dissimilarity assumptions of local functions, and the results are limited to smooth optimization problems. In order to remedy these deficiencies, we develop a novel local dissimilarity invariant convergence theory for \FedProx~and its minibatch stochastic extension through the lens of algorithmic stability. As a result, we contribute to derive several new and deeper insights into \FedProx~for non-convex federated optimization including: 1) convergence guarantees independent on local dissimilarity type conditions; 2) convergence guarantees for non-smooth FL problems; and 3) linear speedup with respect to size of minibatch and number of sampled devices. Our theory for the first time reveals that local dissimilarity and smoothness are not must-have for \FedProx~to get favorable complexity bounds.

## [Human-AI Collaborative Bayesian Optimisation](#)

- Arun Kumar Anjanapura Venkatesh · Santu Rana · Alistair Shilton · Svetha Venkatesh
- abstract@[open-review](#): Abstract Human-AI collaboration looks at harnessing the complementary strengths of both humans and AI. We propose a new method for human-AI collaboration in Bayesian optimisation where the optimum is mainly pursued by the Bayesian optimisation algorithm following complex computation, whilst getting occasional help from the accompanying expert having a deeper knowledge of the underlying physical phenomenon. We expect experts to have some understanding of the correlation structures of the experimental system, but not the location of the optimum. The expert provides feedback by either changing the current recommendation or providing her belief on the good and bad regions of the search space based on the current observations. Our proposed method takes such feedback to build a model that aligns with the expertâ€™s model and then uses it for optimisation. We provide theoretical underpinning on why such an approach may be more efficient than the one without expertâ€™s feedback. The empirical results show the robustness and superiority of our method with promising efficiency gains.

## [MExMI: Pool-based Active Model Extraction Crossover Membership Inference](#)

- Yixin Xiao · Qingqing Ye · Haibo Hu · Huadi Zheng · Chengfang Fang · Jie Shi
- abstract@[open-review](#): With increasing popularity of Machine Learning as a Service (MLaaS), ML models trained from public and proprietary data are deployed in the cloud and deliver prediction services to users. However, as the prediction API becomes a new attack surface, growing concerns have arisen on the confidentiality of ML models. Existing literatures show their vulnerability under model extraction (ME) attacks, while their private training data is vulnerable to another type of attacks, namely, membership inference (MI). In this paper, we show that ME and MI can reinforce each other through a chained and iterative reaction, which can significantly boost ME attack accuracy and improve MI by saving the query cost. As such, we build a framework MExMI for pool-based active model extraction (PAME) to exploit MI through three modules: â€œMI Pre-Filterâ€, â€œMI Post-Filterâ€, and â€œsemi-supervised boostingâ€. Experimental results show that MExMI can improve up to 11.14% from the best known PAME attack and reach 94.07% fidelity with only 16k queries. Furthermore, the precision and recall of the MI attack in MExMI are on par with state-of-the-art MI attack which needs 150k queries.

## [Active Model Adaptation Under Changed Distributions](#)

- Jie-Jing Shao · Lan-Zhe Guo · Xiao-wen Yang · Yu-Feng Li
- abstract@[open-review](#): This work mainly discusses how to make a known model adapt to a variety of changed distributions at a relatively small labeling cost. The technologies inspired by this problem have broad application prospects, especially for non-i.i.d. open-world scenarios. Previous technologies either rely on strong label information or provide no guarantees about generalization performance under varied distribution. This work shows for the first time that the invariance minimization principle could guide active model adaptation to both performance and robustness. We further propose an interactive model adaptation framework, with two sub-modules: active sample selection and invariant relationship learning. Specifically, we formulate the active selection as a mixture distribution separation problem and present an unbiased estimator to address it, which could find the samples that violate the current invariant relationship, with a provable guarantee. The theoretical analysis supports both sub-modules contribute to generalization. A large number of experimental results confirm the promising performance of the new algorithm.

## [Adversarial Task Up-sampling for Meta-learning](#)

- Yichen WU · Long-Kai Huang · Ying Wei
- abstract@[open-review](#): The success of meta-learning on existing benchmarks is predicated on the assumption that the distribution of meta-training tasks covers meta-testing tasks. Frequent violation of the assumption in applications with either insufficient tasks or a very narrow meta-training task distribution leads to memorization or learner overfitting. Recent solutions have pursued augmentation of meta-training tasks, while it is still an open question to generate both correct and sufficiently imaginary tasks. In this paper, we seek an approach that up-samples meta-training tasks from the task representation via a task up-sampling network. Besides, the resulting approach named Adversarial Task Up-sampling (ATU) suffices to generate tasks that can maximally contribute to the latest meta-learner by maximizing an adversarial loss. On few-shot sine regression and image classification datasets, we empirically validate the marked improvement of ATU over state-of-the-art task augmentation strategies in the meta-testing performance and also the quality of up-sampled tasks.

## [Post-hoc estimators for learning to defer to an expert](#)

- Harikrishna Narasimhan · Wittawat Jitkrittum · Aditya Menon · Ankit Rawat · Sanjiv Kumar
- abstract@[open-review](#): Many practical settings allow a learner to defer predictions to one or more costly experts. For example, the learning to defer paradigm allows a learner to defer to a human expert, at some monetary cost. Similarly, the adaptive inference paradigm allows a base model to defer to one or more large models, at some computational cost. The goal in these settings is to learn classification and deferral mechanisms to optimise a suitable accuracy-cost tradeoff. To achieve this, a central issue studied in prior work is the design of a coherent loss function for both mechanisms. In this work, we demonstrate that existing losses have two subtle limitations: they can encourage underfitting when there is a high cost of deferring, and the deferral function can have a weak dependence on the base model predictions. To resolve these issues, we propose a post-hoc training scheme: we train a deferral function on top of a base model, with the objective of predicting to defer when the base model's error probability exceeds the cost of the expert model. This may be viewed as applying a partial surrogate to the ideal deferral loss, which can lead to a tighter approximation and thus better performance. Empirically, we verify the efficacy of post-hoc training on benchmarks for learning to defer and adaptive inference.

## [Explainability Via Causal Self-Talk](#)

- Nicholas Roy · Junkyung Kim · Neil Rabinowitz

- abstract@[open-review](#): Explaining the behavior of AI systems is an important problem that, in practice, is generally avoided. While the XAI community has been developing an abundance of techniques, most incur a set of costs that the wider deep learning community has been unwilling to pay in most situations. We take a pragmatic view of the issue, and define a set of desiderata that capture both the ambitions of XAI and the practical constraints of deep learning. We describe an effective way to satisfy all the desiderata: train the AI system to build a causal model of itself. We develop an instance of this solution for Deep RL agents: Causal Self-Talk. CST operates by training the agent to communicate with itself across time. We implement this method in a simulated 3D environment, and show how it enables agents to generate faithful and semantically-meaningful explanations of their own behavior. Beyond explanations, we also demonstrate that these learned models provide new ways of building semantic control interfaces to AI systems.

## [RTFormer: Efficient Design for Real-Time Semantic Segmentation with Transformer](#)

- Jian Wang · Chenhui Gou · Qiman Wu · Haocheng Feng · Junyu Han · Errui Ding · Jingdong Wang
- abstract@[open-review](#): Recently, transformer-based networks have shown impressive results in semantic segmentation. Yet for real-time semantic segmentation, pure CNN-based approaches still dominate in this field, due to the time-consuming computation mechanism of transformer. We propose RTFormer, an efficient transformer for real-time semantic segmentation, which achieves better trade-off between performance and efficiency than CNN-based models. To achieve high inference efficiency on GPU-like devices, our RTFormer leverages GPU-Friendly Attention with linear complexity and discards the multi-head mechanism. Besides, we find that cross-resolution attention is more efficient to gather global context information for high-resolution branch by spreading the high level knowledge learned from low-resolution branch. Extensive experiments on mainstream benchmarks demonstrate the effectiveness of our proposed RTFormer, it achieves state-of-the-art on Cityscapes and CamVid, and shows promising results on ADE20K.

## [HYPYR: A Hybirdly Normalized Probabilistic Model for Long-Horizon Prediction of Event Sequences](#)

- Siqiao Xue · Xiaoming Shi · James Zhang · Hongyuan Mei
- abstract@[open-review](#): In this paper, we tackle the important yet under-investigated problem of making long-horizon prediction of event sequences. Existing state-of-the-art models do not perform well at this task due to their autoregressive structure. We propose HYPYR, a hybirdly normalized probabilistic model that naturally fits this task: its first part is an autoregressive base model that learns to propose predictions; its second part is an energy function that learns to reweight the proposals such that more realistic predictions end up with higher probabilities. We also propose efficient training and inference algorithms for this model. Experiments on multiple real-world datasets demonstrate that our proposed HYPYR model can significantly outperform previous models at making long-horizon predictions of future events. We also conduct a range of ablation studies to investigate the effectiveness of each component of our proposed methods.

## [Maximum Class Separation as Inductive Bias in One Matrix](#)

- Tejaswi Kasarla · Gertjan Burghouts · Max van Spengler · Elise van der Pol · Rita Cucchiara · Pascal Mettes
- abstract@[open-review](#): Maximizing the separation between classes constitutes a well-known inductive bias in machine learning and a pillar of many traditional algorithms. By default, deep networks are not equipped with this inductive bias and therefore many alternative solutions have been proposed through differential optimization. Current approaches tend to optimize classification and separation jointly: aligning inputs with class vectors and separating class vectors angularly. This paper proposes a simple alternative: encoding maximum separation as an inductive bias in the network by adding one fixed matrix multiplication before computing the softmax activations. The main observation behind our approach is that separation does not require optimization but can be solved in closed-form prior to training and plugged into a network. We outline a recursive approach to obtain the matrix consisting of maximally separable vectors for any number of classes, which can be added with negligible engineering effort and computational overhead. Despite its simple nature, this one matrix multiplication provides real impact. We show that our proposal directly boosts classification, long-tailed recognition, out-of-distribution detection, and open-set recognition, from CIFAR to ImageNet. We find empirically that maximum separation works best as a fixed bias; making the matrix learnable adds nothing to the performance. The closed-form implementation and code to reproduce the experiments are provided in the supplementary materials.

## [HumanLiker: A Human-like Object Detector](#)

- Haoran Wei · Ping Guo · Yangguang Zhu · Chenglong Liu · Peng Wang
- abstract@[open-review](#): Popular object detection models generate bounding boxes in a different way than we humans. As an example, modern detectors yield object box either upon the regression of its center and width/height (center-guided detector), or by grouping paired estimated corners (corner-guided detector). However, that is not the pattern we manually label an object due to high degrees of freedom in searching centers or low efficiency of grouping corners. Empirically, humans run two steps to locate an object bounding box manually: 1) click the mouse at the top-left corner of object, and then drag the mouse to the bottom-right corner; 2) refine the corner positions to make the bounding box more precisely, if necessary. Inspired by this manual labeling process, we propose a novel human-like detector, termed as HumanLiker, which is devised as a two-stage end-to-end detector to simulate the two aforementioned. Like we humans in manual labeling, HumanLiker can effectively avert both the thorny center searching and heuristic corner grouping. Different from the mainstream detector branches, i.e., the center/corner-guided methods, the HumanLiker provides a new paradigm which integrates the advantages of both branches to balance the detection efficiency and bounding box quality. On MS-COCO test-dev set, HumanLiker can achieve 50.2%/51.6% and 53.8%/55.6% in term of AP with ResNeXt-101 and SwinTransformer backbones in single/multi-scale testing, outperforming current popular center/corner-guided baselines (e.g., DETR/CornerNet) by a large margin, with much less training epochs and higher inference FPS. Code will be available soon.

## [An Adaptive Deep RL Method for Non-Stationary Environments with Piecewise Stable Context](#)

- Xiaoyu Chen · Xiangming Zhu · Yufeng Zheng · Pushi Zhang · Li Zhao · Wenzhe Cheng · Peng CHENG · Yongqiang Xiong · Tao Qin · Jianyu Chen · Tie-Yan Liu
- abstract@[open-review](#): One of the key challenges in deploying RL to real-world applications is to adapt to variations of unknown environment contexts, such as changing terrains in robotic tasks and fluctuated bandwidth in congestion control. Existing works on adaptation to unknown environment contexts either assume the contexts are the same for the whole episode or assume the context variables are Markovian. However, in many real-world applications, the environment context usually stays stable for a stochastic period and then changes in an abrupt and unpredictable manner within an episode, resulting in a segment structure, which existing works fail to address. To leverage the segment structure of piecewise stable context in real-world applications, in this paper, we propose a \textit{\textbf{Se}gmented \textbf{C}ontext \textbf{B}elief \textbf{A}ugmented \textbf{D}eep\text{-}\textbf{(SeCBAD)}} RL method. Our method can jointly infer the belief distribution over latent context with the posterior over segment length and perform more accurate belief context inference with observed data within the current context segment. The inferred belief context can be leveraged to augment the state, leading to a policy that can adapt to abrupt variations in context. We demonstrate empirically that SeCBAD can infer context segment length accurately and outperform existing methods on a toy grid world environment and Mujoco tasks with piecewise-stable context.

## [Sparse Probabilistic Circuits via Pruning and Growing](#)

- Meihua Dang · Anji Liu · Guy Van den Broeck
- abstract@[open-review](#): Probabilistic circuits (PCs) are a tractable representation of probability distributions allowing for exact and efficient computation of likelihoods and marginals. There has been significant recent progress on improving the scale and expressiveness of PCs. However, PC training

performance plateaus as model size increases. We discover that most capacity in existing large PC structures is wasted: fully-connected parameter layers are only sparsely used. We propose two operations: pruning and growing, that exploit the sparsity of PC structures. Specifically, the pruning operation removes unimportant sub-networks of the PC for model compression and comes with theoretical guarantees. The growing operation increases model capacity by increasing the dimensions of latent states. By alternately applying pruning and growing, we increase the capacity that is meaningfully used, allowing us to significantly scale up PC learning. Empirically, our learner achieves state-of-the-art likelihoods on MNIST-family image datasets and an Penn Tree Bank language data compared to other PC learners and less tractable deep generative models such as flow-based models and variational autoencoders (VAEs).

## [FiLM: Frequency improved Legendre Memory Model for Long-term Time Series Forecasting](#)

- Tian Zhou · Ziqing MA · xue wang · Qingsong Wen · Liang Sun · Tao Yao · Wotao Yin · Rong Jin
- abstract@[open-review](#): Recent studies have shown that deep learning models such as RNNs and Transformers have brought significant performance gains for long-term forecasting of time series because they effectively utilize historical information. We found, however, that there is still great room for improvement in how to preserve historical information in neural networks while avoiding overfitting to noise present in the history. Addressing this allows better utilization of the capabilities of deep learning models. To this end, we design a \textbf{F}requency \textbf{i}mproved \textbf{L}egendre \textbf{M}emory model, or \textbf{FiLM}: it applies Legendre polynomial projections to approximate historical information, uses Fourier projection to remove noise, and adds a low-rank approximation to speed up computation. Our empirical studies show that the proposed FiLM significantly improves the accuracy of state-of-the-art models in multivariate and univariate long-term forecasting by (\textbf{19.2\%}, \textbf{22.6\%}), respectively. We also demonstrate that the representation module developed in this work can be used as a general plugin to improve the long-term prediction performance of other deep learning modules. Code is available at <https://github.com/tianzhou2011/FiLM/>.

## [Resolving the data ambiguity for periodic crystals](#)

- Daniel Widdowson · Vitaliy Kurlin
- abstract@[open-review](#): The fundamental model of all solid crystalline materials (periodic crystals) is a periodic set of atomic centers considered up to rigid motion in Euclidean space. The major obstacle to materials discovery was highly ambiguous representations that didn't allow fast and reliable comparisons, and led to numerous (near-) duplicates in all experimental databases. This paper introduces the new invariants that are crystal descriptors without false negatives and are called Pointwise Distance Distributions (PDD). The PDD invariants are numerical matrices with a near-linear time complexity and an exactly computable metric. The strongest theoretical result is generic completeness (absence of false positives) for all finite and periodic sets of points in any dimension. The strength of PDD is demonstrated by 200B+ pairwise comparisons of all 660K+ periodic structures from the world's largest Cambridge Structural Database of 1.17M+ known crystals over two days on a modest desktop.

## [Large-Scale Retrieval for Reinforcement Learning](#)

- Peter Humphreys · Arthur Guez · Olivier Tieleman · Laurent Sifre · Theophane Weber · Timothy Lillicrap
- abstract@[open-review](#): Effective decision making involves flexibly relating past experiences and relevant contextual information to a novel situation. In deep reinforcement learning (RL), the dominant paradigm is for an agent to amortise information that helps decision-making into its network weights via gradient descent on training losses. Here, we pursue an alternative approach in which agents can utilise large-scale context-sensitive database lookups to support their parametric computations. This allows agents to directly learn in an end-to-end manner to utilise relevant information to inform their outputs. In addition, new information can be attended to by the agent, without retraining, by simply augmenting the retrieval dataset. We study this approach for offline RL in 9x9 Go, a challenging game for which the vast combinatorial state space privileges generalisation over direct matching to past experiences. We leverage fast, approximate nearest neighbor techniques in order to retrieve relevant data from a set of tens of millions of expert demonstration states. Attending to this information provides a significant boost to prediction accuracy and game-play performance over simply using these demonstrations as training trajectories, providing a compelling demonstration of the value of large-scale retrieval in offline RL agents.

## [Matching in Multi-arm Bandit with Collision](#)

- YiRui Zhang · Siwei Wang · Zhixuan Fang
- abstract@[open-review](#): In this paper, we consider the matching of multi-agent multi-armed bandit problem, i.e., while agents prefer arms with higher expected reward, arms also have preferences on agents. In such case, agents pulling the same arm may encounter collisions, which leads to a reward of zero. For this problem, we design a specific communication protocol which uses deliberate collision to transmit information among agents, and propose a layer-based algorithm that helps establish optimal stable matching between agents and arms. With this subtle communication protocol, our algorithm achieves a state-of-the-art  $\$O(\log T)\$$  regret in the decentralized matching market, and outperforms existing baselines in experimental results.

## [Information bottleneck theory of high-dimensional regression: relevancy, efficiency and optimality](#)

- Vudtiwat Ngampruetikorn · David Schwab
- abstract@[open-review](#): Avoiding overfitting is a central challenge in machine learning, yet many large neural networks readily achieve zero training loss. This puzzling contradiction necessitates new approaches to the study of overfitting. Here we quantify overfitting via residual information, defined as the bits in fitted models that encode noise in training data. Information efficient learning algorithms minimize residual information while maximizing the relevant bits, which are predictive of the unknown generative models. We solve this optimization to obtain the information content of optimal algorithms for a linear regression problem and compare it to that of randomized ridge regression. Our results demonstrate the fundamental trade-off between residual and relevant information and characterize the relative information efficiency of randomized regression with respect to optimal algorithms. Finally, using results from random matrix theory, we reveal the information complexity of learning a linear map in high dimensions and unveil information-theoretic analogs of double and multiple descent phenomena.

## [Sustainable Online Reinforcement Learning for Auto-bidding](#)

- Zhiyu Mou · Yusen Huo · Rongquan Bai · Mingzhou Xie · Chuan Yu · Jian Xu · Bo Zheng
- abstract@[open-review](#): Recently, auto-bidding technique has become an essential tool to increase the return on investment (ROI) for advertisers. Facing the complex and ever-changing bidding environments in the real-world advertising system (RAS), state-of-the-art auto-bidding policies usually leverage reinforcement learning (RL) algorithms to generate real-time bids on behalf of the advertisers. Due to safety concerns, it was believed that the RL process can only be carried out in a virtual advertising system (VAS) which is usually built based on the historical data generated in the RAS. In this paper, we argue that there exists significant gaps between VAS and RAS, making the RL process in the VAS suffer from the problem of inconsistency between online and offline (IBOO). Firstly, we formally define the IBOO and systematically analyze its causes and influences. Then, we design a safety function and prove its Lipschitz smooth property in theory, which provides theoretical foundations for solving the IBOO problem. Moreover, based on this property, we propose a sustainable online RL (SORL) framework to continuously improve the auto-bidding policies through direct interactions with the RAS, which can completely resolve the IBOO problem. Specifically, we devise a safe and efficient online exploration policy that can continuously collect data from the RAS. We also develop a variance-suppressed conservative Q-learning method to improve the auto-bidding policies with the collected data. Extensive experiments on both simulated and real-world advertising systems validate the effectiveness of our approach.

## Dense Interspecies Face Embedding

- Sejong Yang · Subin Jeon · Seonghyeon Nam · Seon Joo Kim
- abstract@[open-review](#): Dense Interspecies Face Embedding (DIFE) is a new direction for understanding faces of various animals by extracting common features among animal faces including human face. There are three main obstacles for interspecies face understanding: (1) lack of animal data compared to human, (2) ambiguous connection between faces of various animals, and (3) extreme shape and style variance. To cope with the lack of data, we utilize multi-teacher knowledge distillation of CSE and StyleGAN2 requiring no additional data or label. Then we synthesize pseudo pair images through the latent space exploration of StyleGAN2 to find implicit associations between different animal faces. Finally, we introduce the semantic matching loss to overcome the problem of extreme shape differences between species. To quantitatively evaluate our method over possible previous methodologies like unsupervised keypoint detection, we perform interspecies facial keypoint transfer on MAFL and AP-10K. Furthermore, the results of other applications like interspecies face image manipulation and dense keypoint transfer are provided.

## Off-Team Learning

- Brandon Cui · Hengyuan Hu · Samuel Sokota · Andrei Lupu · Jakob Foerster
- abstract@[open-review](#): Zero-shot coordination (ZSC) is a coordination framework that evaluates agents under cross-play (XP), i.e. paired with partners independently trained with the same algorithm. Off-belief learning (OBL) is a recent method for ZSC which prevents agents from learning arbitrary conventions. It achieves state-of-the-art results in ZSC, ad-hoc teamplay, and proxy human6 AI coordination in the complex card game Hanabi. However, OBL suffers from two issues: First of all, it relies heavily on a belief model which is trained on a fixed, given policy  $\pi_0$ , but evaluated on the trained policy  $\pi_1$ , potentially leading to large covariate shifts. Secondly, OBL agents are trained only on the narrow distribution induced by their training partner, which means they can easily be taken off distribution at test time. We present two methods addressing these issues. The first, off-team belief learning (OT-BL), is a general algorithm that allows training belief models off-team—i.e., to learn the belief for a given policy,  $\pi$ , on the distribution induced by a different team,  $\pi_b$ . The second, off-team reinforcement learning (OT-RL), allows training a policy off-team—i.e., learning values and policies corresponding to one policy,  $\pi$ , on rollouts from a different team,  $\pi_b$ . We empirically show that OBL+OT-BL outperforms OBL in challenging variants of Hanabi. Furthermore, we show that OBL+OT-RL achieves near-optimal ZSC and can be applied to mitigate large covariate shifts in ad-hoc teamplay and proxy human-AI coordination.

## Learning Consistency-Aware Unsigned Distance Functions Progressively from Raw Point Clouds

- Junsheng Zhou · Baorui Ma · Yu-Shen Liu · Yi Fang · Zhizhong Han
- abstract@[open-review](#): Surface reconstruction for point clouds is an important task in 3D computer vision. Current methods partly address this problem by learning signed distance functions (SDF), which are limited to closed shapes. Some recent methods try to represent non-closed shapes using unsigned distance functions (UDF) but require large scale ground truth distance values during training, which are not trivial to collect through sensors. In this paper, we propose a novel method to learn consistency-aware unsigned distance functions directly from raw point clouds. We achieve this by learning to move 3D queries to reach the surface with a field consistency constrain, where we also enable to progressively estimate a more accurate surface. Specifically, we train a neural network to successive investigate the relationship between 3D queries and the approximated surface by searching for the moving target of queries dynamically to establish a consistent field. Meanwhile, we introduce a polygonization algorithm to extract surfaces directly from the gradient vector field of the learned unsigned distance functions. The experimental results in surface reconstruction for synthetic and real scanned data show significant improvements over the state-of-the-art under the widely used benchmarks.

## Amortized Inference for Causal Structure Learning

- Lars Lorch · Scott Sussex · Jonas Rothfuss · Andreas Krause · Bernhard Schölkopf
- abstract@[open-review](#): Learning causal structure poses a combinatorial search problem that typically involves evaluating structures with a score or independence test. The resulting search is costly, and designing suitable scores or tests that capture prior knowledge is difficult. In this work, we propose to amortize causal structure learning. Rather than searching over structures directly, we train a variational inference model to predict the causal structure from observational or interventional data. This allows us to bypass both the search over graphs and the hand-engineering of suitable score functions. Instead, our inference model acquires domain-specific inductive biases for causal discovery solely from data generated by a simulator. The architecture of our inference model emulates permutation invariances that are crucial for statistical efficiency in structure learning, which facilitates generalization to significantly larger problem instances than seen during training. On synthetic data and semisynthetic gene expression data, our models exhibit robust generalization capabilities when subject to substantial distribution shifts and significantly outperform existing algorithms, especially in the challenging genomics domain.

## Semantic uncertainty intervals for disentangled latent spaces

- Swami Sankaranarayanan · Anastasios Angelopoulos · Stephen Bates · Yaniv Romano · Phillip Isola
- abstract@[open-review](#): Meaningful uncertainty quantification in computer vision requires reasoning about semantic information---say, the hair color of the person in a photo or the location of a car on the street. To this end, recent breakthroughs in generative modeling allow us to represent semantic information in disentangled latent spaces, but providing uncertainties on the semantic latent variables has remained challenging. In this work, we provide principled uncertainty intervals that are guaranteed to contain the true semantic factors for any underlying generative model. The method does the following: (1) it uses quantile regression to output a heuristic uncertainty interval for each element in the latent space (2) calibrates these uncertainties such that they contain the true value of the latent for a new, unseen input. The endpoints of these calibrated intervals can then be propagated through the generator to produce interpretable uncertainty visualizations for each semantic factor. This technique reliably communicates semantically meaningful, principled, and instance-adaptive uncertainty in inverse problems like image super-resolution and image completion.

## Semi-supervised Active Linear Regression

- Nived Rajaraman · Fnu Devvrit · Pranjal Awasthi
- abstract@[open-review](#): Labeled data often comes at a high cost as it may require recruiting human labelers or running costly experiments. At the same time, in many practical scenarios, one already has access to a partially labeled, potentially biased dataset that can help with the learning task at hand. Motivated by such settings, we formally initiate a study of semi-supervised active learning' through the frame of linear regression. Here, the learner has access to a dataset  $\mathcal{X} \in \mathbb{R}^{n_{\text{un}} + n_{\text{lab}}} \times d$  composed of  $n_{\text{un}}$  unlabeled examples that a learner can actively query, and  $n_{\text{lab}}$  labeled a priori. Denoting the true labels by  $\mathcal{Y} \in \mathbb{R}^{n_{\text{un}} + n_{\text{lab}}}$ , the learner's objective is to find  $\widehat{\beta} \in \mathbb{R}^d$  such that  $\|\mathcal{X} \widehat{\beta} - \mathcal{Y}\|_2^2 \leq (1 + \epsilon) \min_{\beta \in \mathbb{R}^d} \|\mathcal{X} \beta - \mathcal{Y}\|_2^2$  while querying the labels of as few unlabeled points as possible. In this paper, we introduce an instance dependent parameter called the reduced rank, denoted  $R_X$ , and propose an efficient algorithm with query complexity  $O(R_X/\epsilon)$ . This result directly implies improved upper bounds for two important special cases: (i) active ridge regression, and (ii) active kernel ridge regression, where the reduced-rank equates to the statistical dimension,  $\text{sd}(\Lambda)$  and "effective dimension",  $d\Lambda$  of the problem respectively, where  $\Lambda \geq 0$  denotes the regularization parameter. Finally, we introduce a distributional version of the problem as a special case of the agnostic formulation we consider earlier; here, for every  $X$ , we prove a matching instance-wise lower bound of  $\Omega(R_X/\epsilon)$  on the query complexity of any algorithm.

## [Active Learning Polynomial Threshold Functions](#)

- Omri Ben-Eliezer · Max Hopkins · Chutong Yang · Hantao Yu
- abstract@[open-review](#): We initiate the study of active learning polynomial threshold functions (PTFs). While traditional lower bounds imply that even univariate quadratics cannot be non-trivially actively learned, we show that allowing the learner basic access to the derivatives of the underlying classifier circumvents this issue and leads to a computationally efficient algorithm for active learning degree-\$d\$ univariate PTFs in \$\tilde{O}((d^3\log(1/\delta))\$ queries. We extend this result to the batch active setting, providing a smooth transition between query complexity and rounds of adaptivity, and also provide near-optimal algorithms for active learning PTFs in several average case settings. Finally, we prove that access to derivatives is insufficient for active learning multivariate PTFs, even those of just two variables.

## [The First Optimal Acceleration of High-Order Methods in Smooth Convex Optimization](#)

- Dmitry Kovalev · Alexander Gasnikov
- abstract@[open-review](#): In this paper, we study the fundamental open question of finding the optimal high-order algorithm for solving smooth convex minimization problems. Arjevani et al. (2019) established the lower bound  $\Omega(\epsilon^{-2/(3p+1)})$  on the number of the  $p$ -th order oracle calls required by an algorithm to find an  $\epsilon$ -accurate solution to the problem, where the  $p$ -th order oracle stands for the computation of the objective function value and the derivatives up to the order  $p$ . However, the existing state-of-the-art high-order methods of Gasnikov et al. (2019b); Bubeck et al. (2019); Jiang et al. (2019) achieve the oracle complexity  $\mathcal{O}(\epsilon^{-2/(3p+1)} \log(1/\epsilon))$ , which does not match the lower bound. The reason for this is that these algorithms require performing a complex binary search procedure, which makes them neither optimal nor practical. We fix this fundamental issue by providing the first algorithm with  $\mathcal{O}(\epsilon^{-2/(3p+1)})$   $p$ -th order oracle complexity.

## [A Policy-Guided Imitation Approach for Offline Reinforcement Learning](#)

- Haoran Xu · Li Jiang · Li Jianxiong · Xianyuan Zhan
- abstract@[open-review](#): Offline reinforcement learning (RL) methods can generally be categorized into two types: RL-based and Imitation-based. RL-based methods could in principle enjoy out-of-distribution generalization but suffer from erroneous off-policy evaluation. Imitation-based methods avoid off-policy evaluation but are too conservative to surpass the dataset. In this study, we propose an alternative approach, inheriting the training stability of imitation-style methods while still allowing logical out-of-distribution generalization. We decompose the conventional reward-maximizing policy in offline RL into a guide-policy and an execute-policy. During training, the guide-policy and execute-policy are learned using only data from the dataset, in a supervised and decoupled manner. During evaluation, the guide-policy guides the execute-policy by telling where it should go so that the reward can be maximized, serving as the \textit{Prophet}. By doing so, our algorithm allows \textit{state-compositionality} from the dataset, rather than \textit{action-compositionality} conducted in prior imitation-style methods. We dumb this new approach Policy-guided Offline RL (\texttt{POR}). \texttt{POR} demonstrates the state-of-the-art performance on D4RL, a standard benchmark for offline RL. We also highlight the benefits of \texttt{POR} in terms of improving with supplementary suboptimal data and easily adapting to new tasks by only changing the guide-policy.

## [Meta Reinforcement Learning with Finite Training Tasks - a Density Estimation Approach](#)

- Zohar Rimon · Aviv Tamar · Gilad Adler
- abstract@[open-review](#): In meta reinforcement learning (meta RL), an agent learns from a set of training tasks how to quickly solve a new task, drawn from the same task distribution. The optimal meta RL policy, a.k.a.~the Bayes-optimal behavior, is well defined, and guarantees optimal reward in expectation, taken with respect to the task distribution. The question we explore in this work is how many training tasks are required to guarantee approximately optimal behavior with high probability. Recent work provided the first such PAC analysis for a model-free setting, where a history-dependent policy was learned from the training tasks. In this work, we propose a different approach: directly learn the task distribution, using density estimation techniques, and then train a policy on the learned task distribution. We show that our approach leads to bounds that depend on the dimension of the task distribution. In particular, in settings where the task distribution lies in a low-dimensional manifold, we extend our analysis to use dimensionality reduction techniques and account for such structure, obtaining significantly better bounds than previous work, which strictly depend on the number of states and actions. The key of our approach is the regularization implied by the kernel density estimation method. We further demonstrate that this regularization is useful in practice, when `plugged in' the state-of-the-art VariBAD meta RL algorithm.

## [Recurrent Video Restoration Transformer with Guided Deformable Attention](#)

- Jingyun Liang · Yuchen Fan · Xiaoyu Xiang · Rakesh Ranjan · Eddy Ilg · Simon Green · Jiezhang Cao · Kai Zhang · Radu Timofte · Luc V Gool
- abstract@[open-review](#): Video restoration aims at restoring multiple high-quality frames from multiple low-quality frames. Existing video restoration methods generally fall into two extreme cases, i.e., they either restore all frames in parallel or restore the video frame by frame in a recurrent way, which would result in different merits and drawbacks. Typically, the former has the advantage of temporal information fusion. However, it suffers from large model size and intensive memory consumption; the latter has a relatively small model size as it shares parameters across frames; however, it lacks long-range dependency modeling ability and parallelizability. In this paper, we attempt to integrate the advantages of the two cases by proposing a recurrent video restoration transformer, namely RVRT. RVRT processes local neighboring frames in parallel within a globally recurrent framework which can achieve a good trade-off between model size, effectiveness, and efficiency. Specifically, RVRT divides the video into multiple clips and uses the previously inferred clip feature to estimate the subsequent clip feature. Within each clip, different frame features are jointly updated with implicit feature aggregation. Across different clips, the guided deformable attention is designed for clip-to-clip alignment, which predicts multiple relevant locations from the whole inferred clip and aggregates their features by the attention mechanism. Extensive experiments on video super-resolution, deblurring, and denoising show that the proposed RVRT achieves state-of-the-art performance on benchmark datasets with balanced model size, testing memory and runtime.

## [Neural Shape Deformation Priors](#)

- Jiapeng Tang · Lev Markhasin · Bi Wang · Justus Thies · Matthias Niessner
- abstract@[open-review](#): We present Neural Shape Deformation Priors, a novel method for shape manipulation that predicts mesh deformations of non-rigid objects from user-provided handle movements. State-of-the-art methods cast this problem as an optimization task, where the input source mesh is iteratively deformed to minimize an objective function according to hand-crafted regularizers such as ARAP. In this work, we learn the deformation behavior based on the underlying geometric properties of a shape, while leveraging a large-scale dataset containing a diverse set of non-rigid deformations. Specifically, given a source mesh and desired target locations of handles that describe the partial surface deformation, we predict a continuous deformation field that is defined in 3D space to describe the space deformation. To this end, we introduce transformer-based deformation networks that represent a shape deformation as a composition of local surface deformations. It learns a set of local latent codes anchored in 3D space, from which we can learn a set of continuous deformation functions for local surfaces. Our method can be applied to challenging deformations and generalizes well to unseen deformations. We validate our approach in experiments using the DeformingThing4D dataset, and compare to both classic optimization-based and recent neural network-based methods.

## [Tiered Reinforcement Learning: Pessimism in the Face of Uncertainty and Constant Regret](#)

- Jiawei Huang · Li Zhao · Tao Qin · Wei Chen · Nan Jiang · Tie-Yan Liu
- abstract@[open-review](#): We propose a new learning framework that captures the tiered structure of many real-world user-interaction applications, where the users can be divided into two groups based on their different tolerance on exploration risks and should be treated separately. In this setting, we simultaneously maintain two policies  $\pi^{\text{O}}$  and  $\pi^{\text{E}}$ :  $\pi^{\text{O}}$  ("foronline") interacts with more risk-tolerant users from the first tier and minimizes regret by balancing exploration and exploitation as usual, while  $\pi^{\text{E}}$  ("forexplorit") exclusively focuses on exploitation for risk-averse users from the second tier utilizing the data collected so far. An important question is whether such a separation yields advantages over the standard online setting (i.e.,  $\pi^{\text{E}} = \pi^{\text{O}}$ ) for the risk-averse users. We individually consider the gap-independent vs.~gap-dependent settings. For the former, we prove that the separation is indeed not beneficial from a minimax perspective. For the latter, we show that if choosing Pessimistic Value Iteration as the exploitation algorithm to produce  $\pi^{\text{E}}$ , we can achieve a constant regret for risk-averse users independent of the number of episodes  $K$ , which is in sharp contrast to the  $\Omega(\log K)$  regret for any online RL algorithms in the same setting, while the regret of  $\pi^{\text{O}}$  (almost) maintains its online regret optimality and does not need to compromise for the success of  $\pi^{\text{E}}$ .

## [The Curse of Unrolling: Rate of Differentiating Through Optimization](#)

- Damien Scieur · Gauthier Gidel · Quentin Bertrand · Fabian Pedregosa
- abstract@[open-review](#): Computing the Jacobian of the solution of an optimization problem is a central problem in machine learning, with applications in hyperparameter optimization, meta-learning, optimization as a layer, and dataset distillation, to name a few. Unrolled differentiation is a popular heuristic that approximates the solution using an iterative solver and differentiates it through the computational path. This work provides a non-asymptotic convergence-rate analysis of this approach on quadratic objectives for gradient descent and the Chebyshev method. We show that to ensure convergence of the Jacobian, we can either 1) choose a large learning rate leading to a fast asymptotic convergence but accept that the algorithm may have an arbitrarily long burn-in phase or 2) choose a smaller learning rate leading to an immediate but slower convergence. We refer to this phenomenon as the curse of unrolling. Finally, we discuss open problems relative to this approach, such as deriving a practical update rule for the optimal unrolling strategy and making novel connections with the field of Sobolev orthogonal polynomials.

## [Density-driven Regularization for Out-of-distribution Detection](#)

- Wenjian Huang · Hao Wang · Jiahao Xia · Chengyan Wang · Jianguo Zhang
- abstract@[open-review](#): Detecting out-of-distribution (OOD) samples is essential for reliably deploying deep learning classifiers in open-world applications. However, existing detectors relying on discriminative probability suffer from the overconfident posterior estimate for OOD data. Other reported approaches either impose strong unproven parametric assumptions to estimate OOD sample density or develop empirical detectors lacking clear theoretical motivations. To address these issues, we propose a theoretical probabilistic framework for OOD detection in deep classification networks, in which two regularization constraints are constructed to reliably estimate sample density to identify OOD. Specifically, the density consistency regularization enforces the agreement between analytical and empirical densities of observable low-dimensional categorical labels. The contrastive distribution regularization separates the densities between in distribution (ID) and distribution-deviated samples. A simple and robust implementation algorithm is also provided, which can be used for any pre-trained neural network classifiers. To the best of our knowledge, we have conducted the most extensive evaluations and comparisons on computer vision benchmarks. The results show that our method significantly outperforms state-of-the-art detectors, and even achieves comparable or better performance than methods utilizing additional large-scale outlier exposure datasets. Our code will be open-sourced upon acceptance.

## [Perturbation Learning Based Anomaly Detection](#)

- Jinyu Cai · Jicong Fan
- abstract@[open-review](#): This paper presents a simple yet effective method for anomaly detection. The main idea is to learn small perturbations to perturb normal data and learn a classifier to classify the normal data and the perturbed data into two different classes. The perturbator and classifier are jointly learned using deep neural networks. Importantly, the perturbations should be as small as possible but the classifier is still able to recognize the perturbed data from unperturbed data. Therefore, the perturbed data are regarded as abnormal data and the classifier provides a decision boundary between the normal data and abnormal data, although the training data do not include any abnormal data. Compared with the state-of-the-art of anomaly detection, our method does not require any assumption about the shape (e.g. hypersphere) of the decision boundary and has fewer hyper-parameters to determine. Empirical studies on benchmark datasets verify the effectiveness and superiority of our method.

## [Neural Stochastic Control](#)

- Jingdong Zhang · Qunxi Zhu · Wei LIN
- abstract@[open-review](#): Control problems are always challenging since they arise from the real-world systems where stochasticity and randomness are of ubiquitous presence. This naturally and urgently calls for developing efficient neural control policies for stabilizing not only the deterministic equations but the stochastic systems as well. Here, in order to meet this paramount call, we propose two types of controllers, viz., the exponential stabilizer (ES) based on the stochastic Lyapunov theory and the asymptotic stabilizer (AS) based on the stochastic asymptotic stability theory. The ES can render the controlled systems exponentially convergent but it requires a long computational time; conversely, the AS makes the training much faster but it can only assure the asymptotic (not the exponential) attractiveness of the control targets. These two stochastic controllers thus are complementary in applications. We also investigate rigorously the linear control in both convergence time and energy cost and numerically compare it with the proposed controllers in these terms. More significantly, we use several representative physical systems to illustrate the usefulness of the proposed controllers in stabilization of dynamical systems.

## [Learning NP-Hard Joint-Assignment planning using GNN: Inference on a Random Graph and Provable Auction-Fitted Q-iteration](#)

- HYUNWOOK KANG · Taehwan Kwon · James R. Morrison · Jinkyoo Park
- abstract@[open-review](#): We develop a theory of inference on a random graph using graph neural networks (GNN) and illustrate its capability to solve NP-hard scheduling problems. We apply the theory to address the challenge of developing a near-optimal learning algorithm to solve the NP-hard problem of scheduling multiple robots/machines with time-varying rewards. In particular, we consider a class of robot/machine scheduling problems called the multi-robot reward collection problem (MRRC). Such MRRC problems well model ride-sharing, pickup-and-delivery, and a variety of related problems. In representing the MRRC problem as a sequential decision-making problem, we observe that each state can be represented as an extension of probabilistic graphical models (PGMs), which we refer to as random PGMs. We then develop a mean-field inference method for random PGMs. We prove that a simple modification of a typical GNN embedding is sufficient to embed a random graph even when the edge presence probabilities are interdependent. We then propose (1) an order-transferable Q-function estimator and (2) an order-transferability-enabled auction to select a joint assignment in polynomial-time. These result in a reinforcement learning framework with at least  $1-1/e$  optimality. Experimental results on solving MRRC problems highlight the near-optimality and transferability of the proposed methods. We also consider minimax multiple traveling salesman problems (minimax-mTSP) and identical parallel machine scheduling problems (IPMS) in the Appendix.

## [LECO: Learnable Episodic Count for Task-Specific Intrinsic Reward](#)

- Daejin Jo · Sungwoong Kim · Daniel Nam · Taehwan Kwon · Seungeun Rho · Jongmin Kim · Donghoon Lee

- abstract@[open-review](#): Episodic count has been widely used to design a simple yet effective intrinsic motivation for reinforcement learning with a sparse reward. However, the use of episodic count in a high-dimensional state space as well as over a long episode time requires a thorough state compression and fast hashing, which hinders rigorous exploitation of it in such hard and complex exploration environments. Moreover, the interference from task-irrelevant observations in the episodic count may cause its intrinsic motivation to overlook task-related important changes of states, and the novelty in an episodic manner can lead to repeatedly revisit the familiar states across episodes. In order to resolve these issues, in this paper, we propose a learnable hash-based episodic count, which we name LECO, that efficiently performs as a task-specific intrinsic reward in hard exploration problems. In particular, the proposed intrinsic reward consists of the episodic novelty and the task-specific modulation where the former employs a vector quantized variational autoencoder to automatically obtain the discrete state codes for fast counting while the latter regulates the episodic novelty by learning a modulator to optimize the task-specific extrinsic reward. The proposed LECO specifically enables the automatic transition from exploration to exploitation during reinforcement learning. We experimentally show that in contrast to the previous exploration methods LECO successfully solves hard exploration problems and also scales to large state spaces through the most difficult tasks in MiniGrid and DMLab environments.

## [What Can Transformers Learn In-Context? A Case Study of Simple Function Classes](#)

- Shivam Garg · Dimitris Tsipras · Gregory Valiant · Percy Liang
- abstract@[open-review](#): In-context learning is the ability of a model to condition on a prompt sequence consisting of in-context examples (input-output pairs corresponding to some task) along with a new query input, and generate the corresponding output. Crucially, in-context learning happens only at inference time without any parameter updates to the model. While large language models such as GPT-3 exhibit some ability to perform in-context learning, it is unclear what the relationship is between tasks on which this succeeds and what is present in the training data. To investigate this, we consider the problem of training a model to in-context learn a function class (e.g., linear functions): given data derived from some functions in the class, can we train a model (e.g., a Transformer) to in-context learn most functions from that class? We show empirically that standard Transformers can be trained from scratch to perform in-context learning of linear functions—that is, the trained model is able to learn unseen linear functions from in-context examples with performance comparable to the optimal least squares estimator. In fact, in-context learning is possible even under two forms of distribution shift: (i) between the training data of the Transformer and inference-time prompts, and (ii) between the in-context examples and the query input during inference. We also show that we can train Transformers to in-context learn more complex function classes: sparse linear functions where the model outperforms least squares and nearly matches the performance of Lasso, and two-layer neural networks where the model performs comparably to neural networks trained on in-context examples using gradient descent.

## [Conditional Independence Testing with Heteroskedastic Data and Applications to Causal Discovery](#)

- Wiebke Gänther · Urmi Ninad · Jonas Wahl · Jakob Runge
- abstract@[open-review](#): Conditional independence (CI) testing is frequently used in data analysis and machine learning for various scientific fields and it forms the basis of constraint-based causal discovery. Oftentimes, CI testing relies on strong, rather unrealistic assumptions. One of these assumptions is homoskedasticity, in other words, a constant conditional variance is assumed. We frame heteroskedasticity in a structural causal model framework and present an adaptation of the partial correlation CI test that works well in the presence of heteroskedastic noise, given that expert knowledge about the heteroskedastic relationships is available. Further, we provide theoretical consistency results for the proposed CI test which carry over to causal discovery under certain assumptions. Numerical causal discovery experiments demonstrate that the adapted partial correlation CI test outperforms the standard test in the presence of heteroskedasticity and is on par for the homoskedastic case. Finally, we discuss the general challenges and limits as to how expert knowledge about heteroskedasticity can be accounted for in causal discovery.

## [A Unified Sequence Interface for Vision Tasks](#)

- Ting Chen · Saurabh Saxena · Lala Li · Tsung-Yi Lin · David Fleet · Geoffrey E Hinton
- abstract@[open-review](#): While language tasks are naturally expressed in a single, unified, modeling framework, i.e., generating sequences of tokens, this has not been the case in computer vision. As a result, there is a proliferation of distinct architectures and loss functions for different vision tasks. In this work we show that a diverse set of ``core'' computer vision tasks can also be unified if formulated in terms of a shared pixel-to-sequence interface. We focus on four tasks, namely, object detection, instance segmentation, keypoint detection, and image captioning, all with diverse types of outputs, e.g., bounding boxes or dense masks. Despite that, by formulating the output of each task as a sequence of discrete tokens with a unified interface, we show that one can train a neural network with a single model architecture and loss function on all these tasks, with no task-specific customization. To solve a specific task, we use a short prompt as task description, and the sequence output adapts to the prompt so it can produce task-specific output. We show that such a model can achieve competitive performance compared to well-established task-specific models.

## [Generalized Laplacian Eigenmaps](#)

- Hao Zhu · Piotr Koniusz
- abstract@[open-review](#): Graph contrastive learning attracts/disperses node representations for similar/dissimilar node pairs under some notion of similarity. It may be combined with a low-dimensional embedding of nodes to preserve intrinsic and structural properties of a graph. Some recent graph contrastive methods combine traditional graph embedding and negative sampling into one framework, which minimizes the trace difference between the within-class scatter matrix encapsulating the graph connectivity and the total scatter matrix encapsulating negative sampling. In this paper, we propose a more essential framework for graph embedding, called Generalized Laplacian Eigenmaps (GLEN), to learn graph representation by maximizing the rank difference between the total scatter matrix and the within-class scatter matrix, resulting in the minimum class separation guarantee. However, the rank difference minimization is an NP-hard problem. Herein, we replace the trace difference that corresponds to the difference of nuclear norms by the difference of logdet expressions, which we argue is a more accurate surrogate for the NP-hard rank difference than the trace difference. We show that the logdet loss can be interpreted as an upper bound of the Jensen-Bregman LogDet Divergence (JBLD), and the Affine-invariant Riemannian metric (AIRM) while enjoying a lesser computational burden. We show on popular benchmarks/backbones that GLEN offers favourable accuracy/scalability compared to state-of-the-art baselines.

## [TabNAS: Rejection Sampling for Neural Architecture Search on Tabular Datasets](#)

- Chengrun Yang · Gabriel Bender · Hanxiao Liu · Pieter-Jan Kindermans · Madeleine Udell · Yifeng Lu · Quoc V Le · Da Huang
- abstract@[open-review](#): The best neural architecture for a given machine learning problem depends on many factors: not only the complexity and structure of the dataset, but also on resource constraints including latency, compute, energy consumption, etc. Neural architecture search (NAS) for tabular datasets is an important but under-explored problem. Previous NAS algorithms designed for image search spaces incorporate resource constraints directly into the reinforcement learning (RL) rewards. However, for NAS on tabular datasets, this protocol often discovers suboptimal architectures. This paper develops TabNAS, a new and more effective approach to handle resource constraints in tabular NAS using an RL controller motivated by the idea of rejection sampling. TabNAS immediately discards any architecture that violates the resource constraints without training or learning from that architecture. TabNAS uses a Monte-Carlo-based correction to the RL policy gradient update to account for this extra filtering step. Results on several tabular datasets demonstrate the superiority of TabNAS over previous reward-shaping methods: it finds better models that obey the constraints.

## [FedPop: A Bayesian Approach for Personalised Federated Learning](#)

- Nikita Kotelevskii · Maxime Vono · Alain Durmus · Eric Moulines

- abstract@[open-review](#): Personalised federated learning (FL) approaches aim at collaboratively learning a machine learning model tailored for each client. Albeit promising advances have been made in this direction, most of existing personalised FL works do not allow for uncertainty quantification which is crucial in many applications. In addition, personalisation in the cross-device setting still involves important issues, especially for new clients or those having small data sets. This paper aims at filling this gap. To this end, we propose a novel methodology coined FedPop by recasting personalised FL into the population modeling paradigm where clients' models involve fixed common population parameters and random individual ones, aiming at explaining data heterogeneity. To derive convergence guarantees for our scheme, we introduce a new class of federated stochastic optimisation algorithms which relies on Markov chain Monte Carlo methods. Compared to existing personalised FL methods, the proposed methodology has important benefits: it is robust to client drift, practical for inference on new clients, and above all, enables uncertainty quantification under mild computational and memory overheads. We provide non-asymptotic convergence guarantees for the proposed algorithms and illustrate their performances on various personalised federated learning tasks.

## [Video Diffusion Models](#)

- Jonathan Ho · Tim Salimans · Alexey Gritsenko · William Chan · Mohammad Norouzi · David Fleet
- abstract@[open-review](#): Generating temporally coherent high fidelity video is an important milestone in generative modeling research. We make progress towards this milestone by proposing a diffusion model for video generation that shows very promising initial results. Our model is a natural extension of the standard image diffusion architecture, and it enables jointly training from image and video data, which we find to reduce the variance of minibatch gradients and speed up optimization. To generate long and higher resolution videos we introduce a new conditional sampling technique for spatial and temporal video extension that performs better than previously proposed methods. We present the first results on a large text-conditioned video generation task, as well as state-of-the-art results on established benchmarks for video prediction and unconditional video generation.

## [Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding](#)

- Chitwan Saharia · William Chan · Saurabh Saxena · Lala Li · Jay Whang · Emily Denton · Seyed Kamyar Seyed Ghasemipour · Raphael Gontijo Lopes · Burcu Karagol Ayan · Tim Salimans · Jonathan Ho · David Fleet · Mohammad Norouzi
- abstract@[open-review](#): We present Imagen, a text-to-image diffusion model with an unprecedented degree of photorealism and a deep level of language understanding. Imagen builds on the power of large transformer language models in understanding text and hinges on the strength of diffusion models in high-fidelity image generation. Our key discovery is that generic large language models (e.g., T5), pretrained on text-only corpora, are surprisingly effective at encoding text for image synthesis: increasing the size of the language model in Imagen boosts both sample fidelity and image-text alignment much more than increasing the size of the image diffusion model. Imagen achieves a new state-of-the-art FID score of 7.27 on the COCO dataset, without ever training on COCO, and human raters find Imagen samples to be on par with the COCO data itself in image-text alignment. To assess text-to-image models in greater depth, we introduce DrawBench, a comprehensive and challenging benchmark for text-to-image models. With DrawBench, we compare Imagen with recent methods including VQ-GAN+CLIP, Latent Diffusion Models, and DALL-E 2, and find that human raters prefer Imagen over other models in side-by-side comparisons, both in terms of sample quality and image-text alignment.

## [DPM-Solver: A Fast ODE Solver for Diffusion Probabilistic Model Sampling in Around 10 Steps](#)

- Cheng Lu · Yuhao Zhou · Fan Bao · Jianfei Chen · Chongxuan LI · Jun Zhu
- abstract@[open-review](#): Diffusion probabilistic models (DPMs) are emerging powerful generative models. Despite their high-quality generation performance, DPMs still suffer from their slow sampling as they generally need hundreds or thousands of sequential function evaluations (steps) of large neural networks to draw a sample. Sampling from DPMs can be viewed alternatively as solving the corresponding diffusion ordinary differential equations (ODEs). In this work, we propose an exact formulation of the solution of diffusion ODEs. The formulation analytically computes the linear part of the solution, rather than leaving all terms to black-box ODE solvers as adopted in previous works. By applying change-of-variable, the solution can be equivalently simplified to an exponentially weighted integral of the neural network. Based on our formulation, we propose DPM-Solver, a fast dedicated high-order solver for diffusion ODEs with the convergence order guarantee. DPM-Solver is suitable for both discrete-time and continuous-time DPMs without any further training. Experimental results show that DPM-Solver can generate high-quality samples in only 10 to 20 function evaluations on various datasets. We achieve 4.70 FID in 10 function evaluations and 2.87 FID in 20 function evaluations on the CIFAR10 dataset, and a 4~16x speedup compared with previous state-of-the-art training-free samplers on various datasets.

## [Coded Residual Transform for Generalizable Deep Metric Learning](#)

- Shichao Kan · Yixiong Liang · Min Li · Yigang Cen · Jianxin Wang · Zhihai He
- abstract@[open-review](#): A fundamental challenge in deep metric learning is the generalization capability of the feature embedding network model since the embedding network learned on training classes need to be evaluated on new test classes. To address this challenge, in this paper, we introduce a new method called coded residual transform (CRT) for deep metric learning to significantly improve its generalization capability. Specifically, we learn a set of diversified prototype features, project the feature map onto each prototype, and then encode its features using their projection residuals weighted by their correlation coefficients with each prototype. The proposed CRT method has the following two unique characteristics. First, it represents and encodes the feature map from a set of complimentary perspectives based on projections onto diversified prototypes. Second, unlike existing transformer-based feature representation approaches which encode the original values of features based on global correlation analysis, the proposed coded residual transform encodes the relative differences between the original features and their projected prototypes. Embedding space density and spectral decay analysis show that this multi perspective projection onto diversified prototypes and coded residual representation are able to achieve significantly improved generalization capability in metric learning. Finally, to further enhance the generalization performance, we propose to enforce the consistency on their feature similarity matrices between coded residual transforms with different sizes of projection prototypes and embedding dimensions. Our extensive experimental results and ablation studies demonstrate that the proposed CRT method outperform the state-of-the-art deep metric learning methods by large margins and improving upon the current best method by up to 4.28% on the CUB dataset.

## [Masked Autoencoding for Scalable and Generalizable Decision Making](#)

- Fangchen Liu · Hao Liu · Aditya Grover · Pieter Abbeel
- abstract@[open-review](#): We are interested in learning scalable agents for reinforcement learning that can learn from large-scale, diverse-quality sequential data similar to current large vision and language models. To this end, this paper presents masked decision prediction (MaskDP), a simple and scalable self-supervised pretraining method for reinforcement learning (RL), and behavioral cloning (BC). In our MaskDP approach, we employ a masked autoencoder (MAE) to state-action trajectories, wherein we randomly mask state and action tokens and reconstruct the missing data. By doing so, the model is required to infer masked out states and actions and extract information about dynamics. We find that masking different proportions of the input sequence significantly helps with learning a better model that generalizes well to multiple downstream tasks. In our empirical study, we find that a MaskDP model gains the capability of zero-shot transfer to new BC tasks, such as single and multiple goal-reaching, and it can zero-shot infer skills from a few example transitions. In addition, MaskDP transfers well to offline RL and shows promising scaling behavior w.r.t. to model size. It is amenable to data-efficient finetuning, achieving competitive results with prior methods based on autoregressive pre-training.

## [Time Dimension Dances with Simplicial Complexes: Zigzag Filtration Curve based Supra-Hodge Convolution Networks for Time-series Forecasting](#)

- Yuzhou Chen · Yulia Gel · H. Vincent Poor
- abstract@[open-review](#): Graph neural networks (GNN) offer a new powerful alternative for multivariate time series forecasting, demonstrating remarkable success in a variety of spatio-temporal applications, from urban flow monitoring systems to health care informatics to financial analytics. Yet, such GNN models pre-dominantly capture only lower order interactions, that is, pairwise relations among nodes, and also largely ignore intrinsic time-conditioned information on the underlying topology of multivariate time series. To address these limitations, we propose a new time-aware GNN architecture which amplifies the power of the recently emerged simplicial neural networks with a time-conditioned topological knowledge representation in a form of zigzag persistence. That is, our new approach, Zigzag Filtration Curve based Supra-Hodge Convolution Networks (ZFC-SHCN) is built upon the two main components: (i) a new highly computationally efficient zigzag persistence curve which allows us to systematically encode time-conditioned topological information, and (ii) a new temporal multiplex graph representation module for learning higher-order network interactions. We discuss theoretical properties of the proposed time-conditioned topological knowledge representation and extensively validate the new time-aware ZFC-SHCN model in conjunction with time series forecasting on a broad range of synthetic and real world datasets: traffic flows, COVID-19 biosurveillance, Ethereum blockchain, surface air temperature, and vector autoregressions. Our experiments demonstrate that ZFC-SHCN achieves the state-of-the-art performance with lower requirements on computational costs.

## [Error Correction Code Transformer](#)

- Yoni Choukroun · Lior Wolf
- abstract@[open-review](#): Error correction code is a major part of the physical communication layer, ensuring the reliable transfer of data over noisy channels. Recently, neural decoders were shown to outperform classical decoding techniques. However, the existing neural approaches present strong overfitting, due to the exponential training complexity, or a restrictive inductive bias, due to reliance on Belief Propagation. Recently, Transformers have become methods of choice in many applications, thanks to their ability to represent complex interactions between elements. In this work, we propose to extend for the first time the Transformer architecture to the soft decoding of linear codes at arbitrary block lengths. We encode each channel's output dimension to a high dimension for better representation of the bits' information to be processed separately. The element-wise processing allows the analysis of channel output reliability, while the algebraic code and the interaction between the bits are inserted into the model via an adapted masked self-attention module. The proposed approach demonstrates the power and flexibility of Transformers and outperforms existing state-of-the-art neural decoders by large margins, at a fraction of their time complexity.

## [A Unified Analysis of Federated Learning with Arbitrary Client Participation](#)

- Shiqiang Wang · Mingyue Ji
- abstract@[open-review](#): Federated learning (FL) faces challenges of intermittent client availability and computation/communication efficiency. As a result, only a small subset of clients can participate in FL at a given time. It is important to understand how partial client participation affects convergence, but most existing works have either considered idealized participation patterns or obtained results with non-zero optimality error for generic patterns. In this paper, we provide a unified convergence analysis for FL with arbitrary client participation. We first introduce a generalized version of federated averaging (FedAvg) that amplifies parameter updates at an interval of multiple FL rounds. Then, we present a novel analysis that captures the effect of client participation in a single term. By analyzing this term, we obtain convergence upper bounds for a wide range of participation patterns, including both non-stochastic and stochastic cases, which match either the lower bound of stochastic gradient descent (SGD) or the state-of-the-art results in specific settings. We also discuss various insights, recommendations, and experimental results.

## [Fault-Aware Neural Code Rankers](#)

- Jeevana Priya Inala · Chenglong Wang · Mei Yang · Andres Codas · Mark Encarnaciñ · Shuvendu Lahiri · Madanlal Musuvathi · Jianfeng Gao
- abstract@[open-review](#): Large language models (LLMs) have demonstrated an impressive ability to generate code for various programming tasks. In many instances, LLMs can generate a correct program for a task when given numerous trials. Consequently, a recent trend is to do large scale sampling of programs using a model and then filtering/ranking the programs based on the program execution on a small number of known unit tests to select one candidate solution. However, these approaches assume that the unit tests are given and assume the ability to safely execute the generated programs (which can do arbitrary dangerous operations such as file manipulations). Both of the above assumptions are impractical in real-world software development. In this paper, we propose fault-aware neural code rankers that can predict the correctness of a sampled program without executing it. The fault-aware rankers are trained to predict different kinds of execution information such as predicting the exact compile/runtime error type (e.g., an IndexError or a TypeError). We show that our fault-aware rankers can significantly increase the pass@1 accuracy of various code generation models (including Codex, GPT-Neo, GPT-J) on APPS, HumanEval and MBPP datasets.

## [Learning to Find Proofs and Theorems by Learning to Refine Search Strategies](#)

- Jonathan Laurent · András Platzer
- abstract@[open-review](#): We propose a new approach to automated theorem proving and deductive program synthesis where an AlphaZero-style agent is self-training to refine a high-level expert strategy expressed as a nondeterministic program. An analogous teacher agent is self-training to generate tasks of suitable relevance and difficulty for the learner. This allows leveraging minimal amounts of domain knowledge to tackle problems for which training data is unavailable or hard to synthesize. We illustrate our approach on the problem of loop invariant synthesis for imperative programs and using neural networks to refine both the teacher and solver strategies.

## [Interventions, Where and How? Bayesian Active Causal Discovery at Scale](#)

- Panagiotis Tigas · Yashas Annadani · Andrew Jesson · Bernhard Schölkopf · Yarin Gal · Stefan Bauer
- abstract@[open-review](#): Causal discovery from observational and interventional data is challenging due to limited data and non-identifiability which introduces uncertainties in estimating the underlying structural causal model (SCM). Incorporating these uncertainties and selecting optimal experiments (interventions) to perform can help to identify the true SCM faster. Existing methods in experimental design for causal discovery from limited data either rely on linear assumptions for the SCM or select only the intervention target. In this paper, we incorporate recent advances in Bayesian causal discovery into the Bayesian optimal experimental design framework, which allows for active causal discovery of nonlinear, large SCMs, while selecting both the target and the value to intervene with. We demonstrate the performance of the proposed method on synthetic graphs (Erdos-Rényi, Scale Free) for both linear and nonlinear SCMs as well as on the \textit{in-silico} single-cell gene regulatory network dataset, DREAM.

## [Flexible Neural Image Compression via Code Editing](#)

- Chenjian Gao · Tongda Xu · Dailan He · Yan Wang · Hongwei Qin
- abstract@[open-review](#): Neural image compression (NIC) has outperformed traditional image codecs in rate-distortion (R-D) performance. However, it usually requires a dedicated encoder-decoder pair for each point on R-D curve, which greatly hinders its practical deployment. While some recent works have enabled bitrate control via conditional coding, they impose strong prior during training and provide limited flexibility. In this paper we propose Code Editing, a highly flexible coding method for NIC based on semi-amortized inference. And we further improve our Code Editing via adaptive quantization. We demonstrate our method under various conditions and we show that our code editing surpass existing variable-rate methods through experiment. Our

work is a new paradigm for variable bitrate NIC, and it is the first to achieve continuous bitrate control, ROI coding and multi-distortion trade-off with a single decoder.

## [Dual-Curriculum Contrastive Multi-Instance Learning for Cancer Prognosis Analysis with Whole Slide Images](#)

- CHAO TU · YU ZHANG · Zhenyuan Ning
- abstract@[open-review](#): The multi-instance learning (MIL) has advanced cancer prognosis analysis with whole slide images (WSIs). However, current MIL methods for WSI analysis still confront unique challenges. Previous methods typically generate instance representations via a pre-trained model or a model trained by the instances with bag-level annotations, which, however, may not generalize well to the downstream task due to the introduction of excessive label noises and the lack of fine-grained information across multi-magnification WSIs. Additionally, existing methods generally aggregate instance representations as bag ones for prognosis prediction and have no consideration of intra-bag redundancy and inter-bag discrimination. To address these issues, we propose a dual-curriculum contrastive MIL method for cancer prognosis analysis with WSIs. The proposed method consists of two curriculums, i.e., saliency-guided weakly-supervised instance encoding with cross-scale tiles and contrastive-enhanced soft-bag prognosis inference. Extensive experiments on three public datasets demonstrate that our method outperforms state-of-the-art methods in this field. The code is available.

## [Redundant representations help generalization in wide neural networks](#)

- Diego Doimo · Aldo Glielmo · Sebastian Goldt · Alessandro Laio
- abstract@[open-review](#): Deep neural networks (DNNs) defy the classical bias-variance trade-off: adding parameters to a DNN that interpolates its training data will typically improve its generalization performance. Explaining the mechanism behind this ``benign overfitting'' in deep networks remains an outstanding challenge. Here, we study the last hidden layer representations of various state-of-the-art convolutional neural networks and find that if the last hidden representation is wide enough, its neurons tend to split into groups which carry identical information, and differ from each other only by a statistically independent noise. The number of such groups increases linearly with the width of the layer, but only if the width is above a critical value. We show that redundant neurons appear only when training process reaches interpolation and the training error is zero.

## [Shadow Knowledge Distillation: Bridging Offline and Online Knowledge Transfer](#)

- Lujun Li · ZHE JIN
- abstract@[open-review](#): Knowledge distillation can be generally divided into offline and online categories according to whether teacher model is pre-trained and persistent during the distillation process. Offline distillation can employ existing models yet always demonstrates inferior performance than online ones. In this paper, we first empirically show that the essential factor for their performance gap lies in the reversed distillation from student to teacher, rather than the training fashion. Offline distillation can achieve competitive performance gain by fine-tuning pre-trained teacher to adapt student with such reversed distillation. However, this fine-tuning process still costs lots of training budgets. To alleviate this dilemma, we propose SHAKE, a simple yet effective SHAdow KnowlEdge transfer framework to bridge offline and online distillation, which trades the accuracy with efficiency. Specifically, we build an extra shadow head on the student's backbone to mimic the predictions of pre-trained teacher as its shadow. Then, this shadow head is leveraged as a proxy teacher to perform bidirectional distillation with student on the fly. In this way, SHAKE not only updates this student-aware proxy teacher with the knowledge of pre-trained model, but also greatly optimizes costs of augmented reversed distillation. Extensive experiments on classification and object detection tasks demonstrate that our technique achieves state-of-the-art results with different CNNs and Vision Transformer models. Additionally, our method shows strong compatibility with multi-teacher and augmentation strategies by gaining additional performance improvement. Code will be made publicly available.

## [Challenging Common Assumptions in Convex Reinforcement Learning](#)

- Mirco Mutti · Riccardo De Santi · Piersilvio De Bartolomeis · Marcello Restelli
- abstract@[open-review](#): The classic Reinforcement Learning (RL) formulation concerns the maximization of a scalar reward function. More recently, convex RL has been introduced to extend the RL formulation to all the objectives that are convex functions of the state distribution induced by a policy. Notably, convex RL covers several relevant applications that do not fall into the scalar formulation, including imitation learning, risk-averse RL, and pure exploration. In classic RL, it is common to optimize an infinite trials objective, which accounts for the state distribution instead of the empirical state visitation frequencies, even though the actual number of trajectories is always finite in practice. This is theoretically sound since the infinite trials and finite trials objectives are equivalent and thus lead to the same optimal policy. In this paper, we show that this hidden assumption does not hold in convex RL. In particular, we prove that erroneously optimizing the infinite trials objective in place of the actual finite trials one, as it is usually done, can lead to a significant approximation error. Since the finite trials setting is the default in both simulated and real-world RL, we believe shedding light on this issue will lead to better approaches and methodologies for convex RL, impacting relevant research areas such as imitation learning, risk-averse RL, and pure exploration among others.

## [Co-Modality Imbalanced Graph Contrastive Learning](#)

- Yiyue Qian · Chunhui Zhang · Yiming Zhang · Qianlong Wen · Yanfang Ye · Chuxu Zhang
- abstract@[open-review](#): Graph contrastive learning (GCL), leveraging graph augmentations to convert graphs into different views and further train graph neural networks (GNNs), has achieved considerable success on graph benchmark datasets. Yet, there are still some gaps in directly applying existing GCL methods to real-world data. First, handcrafted graph augmentations require expert knowledge as well as trials and errors, but still can not yield consistent performance on multiple tasks. Second, most real-world graph data present imbalanced distribution but existing GCL methods are not immune to data imbalance. Therefore, this work proposes to explicitly tackle these challenges, via a principled framework called \textit{Co-Modality Imbalanced Graph Contratsive Learning with Network Pruning} (\textbf{CMI-GCL}) to automatically generate contrastive pairs without expert knowledge and further learn balanced representation over unlabeled data. Specifically, we design inter-modality GCL to automatically generate contrastive pairs (e.g., node-image) based on node content. Inspired by the fact that minority samples can be ``forgotten'' by pruning deep neural networks, we naturally extend the ad-hoc compression technique, network pruning, to our GCL framework for detecting minority nodes. Based on this, we co-train two pruned encoders (e.g., GNN and image encoder) in different modalities by pushing the corresponding node-image pairs together and irrelevant node-image pairs away. Meanwhile, we also propose intra-modality GCL by co-training non-pruned GNN and pruned GNN, to ensure node embeddings with similar attributed features stay closed. By applying pre-trained CMI-GCL to two fine-tuning modes, we demonstrate that our model significantly outperforms state-of-the-art baseline models and learns more balanced representations on real-world graph datasets.

## [Learnable Polyphase Sampling for Shift Invariant and Equivariant Convolutional Networks](#)

- Renan A. Rojas-Gomez · TeckYian Lim · Alex Schwing · Minh Do · Raymond A. Yeh
- abstract@[open-review](#): We propose learnable polyphase sampling (LPS), a pair of learnable down/upsampling layers that enable truly shift-invariant and equivariant convolutional networks. LPS can be trained end-to-end from data and generalizes existing handcrafted downsampling layers. It is widely applicable as it can be integrated into any convolutional network by replacing down/upsampling layers. We evaluate LPS on image classification and semantic segmentation. Experiments show that LPS is on-par with or outperforms existing methods in both performance and shift consistency. For the first time, we achieve true shift-equivariance on semantic segmentation (PASCAL VOC), i.e., 100% shift consistency, outperforming baselines by an absolute 3.3%.

## Self-Supervised Learning Through Efference Copies

- Franz Scherr · Qinghai Guo · Timoleon Moraits
- abstract@[open-review](#): State-of-the-art (SOTA) machine learning (ML) without labels is based on self-supervised learning (SSL). SSL advances are highly demanded given the scarcity of labelled data and the cost of human supervision. Biology can offer inspiration for advanced SSL models that could also bring insights into learning in the brain. SSL commonly transforms each training datapoint into a pair of views, uses the knowledge of this pairing as a positive self-supervisory sign, and potentially opposes it to unrelated, negative examples. Here, first we show that this type of self-supervision is a concrete -- albeit constrained -- implementation of a concept from neuroscience: the Efference Copy (EC). Specifically, like SSL, the brain also transforms the environment through efference, i.e. outgoing motor commands, however it sends to itself an EC of the full commands, i.e. more than a mere sign. Second, we implement this insight as an SSL framework of Self-supervision Through Efference Copies (S-TEC), and we show empirically that S-TEC improves recent strong SSL baselines on image classification. Third, we hypothesize and provide preliminary evidence that S-TEC's improvement to the representation is by placing towards the border of each class heavily transformed data specifically, resulting in better separation between the classes' supports and more structured inter-class space. Overall, S-TEC conceptually generalizes the SSL framework, and hypothesizes a testable positive influence from motor outputs onto sensory representations of the brain.

## On the Generalization Power of the Overfitted Three-Layer Neural Tangent Kernel Model

- Peizhong Ju · Xiaojun Lin · Ness Shroff
- abstract@[open-review](#): In this paper, we study the generalization performance of overparameterized 3-layer NTK models. We show that, for a specific set of ground-truth functions (which we refer to as the "learnable set"), the test error of the overfitted 3-layer NTK is upper bounded by an expression that decreases with the number of neurons of the two hidden layers. Different from 2-layer NTK where there exists only one hidden-layer, the 3-layer NTK involves interactions between two hidden-layers. Our upper bound reveals that, between the two hidden-layers, the test error descends faster with respect to the number of neurons in the second hidden-layer (the one closer to the output) than with respect to that in the first hidden-layer (the one closer to the input). We also show that the learnable set of 3-layer NTK without bias is no smaller than that of 2-layer NTK models with various choices of bias in the neurons. However, in terms of the actual generalization performance, our results suggest that 3-layer NTK is much less sensitive to the choices of bias than 2-layer NTK, especially when the input dimension is large.

## HUMUS-Net: Hybrid Unrolled Multi-scale Network Architecture for Accelerated MRI Reconstruction

- Zalan Fabian · Berk Tinaz · Mahdi Soltanolkotabi
- abstract@[open-review](#): In accelerated MRI reconstruction, the anatomy of a patient is recovered from a set of undersampled and noisy measurements. Deep learning approaches have been proven to be successful in solving this ill-posed inverse problem and are capable of producing very high quality reconstructions. However, current architectures heavily rely on convolutions, that are content-independent and have difficulties modeling long-range dependencies in images. Recently, Transformers, the workhorse of contemporary natural language processing, have emerged as powerful building blocks for a multitude of vision tasks. These models split input images into non-overlapping patches, embed the patches into lower-dimensional tokens and utilize a self-attention mechanism that does not suffer from the aforementioned weaknesses of convolutional architectures. However, Transformers incur extremely high compute and memory cost when 1) the input image resolution is high and 2) when the image needs to be split into a large number of patches to preserve fine detail information, both of which are typical in low-level vision problems such as MRI reconstruction, having a compounding effect. To tackle these challenges, we propose HUMUS-Net, a hybrid architecture that combines the beneficial implicit bias and efficiency of convolutions with the power of Transformer blocks in an unrolled and multi-scale network. HUMUS-Net extracts high-resolution features via convolutional blocks and refines low-resolution features via a novel Transformer-based multi-scale feature extractor. Features from both levels are then synthesized into a high-resolution output reconstruction. Our network establishes new state of the art on the largest publicly available MRI dataset, the fastMRI dataset. We further demonstrate the performance of HUMUS-Net on two other popular MRI datasets and perform fine-grained ablation studies to validate our design.

## Minimax Regret for Cascading Bandits

- Daniel Vial · Sujay Sanghavi · Sanjay Shakkottai · R. Srikant
- abstract@[open-review](#): Cascading bandits is a natural and popular model that frames the task of learning to rank from Bernoulli click feedback in a bandit setting. For the case of unstructured rewards, we prove matching upper and lower bounds for the problem-independent (i.e., gap-free) regret, both of which strictly improve the best known. A key observation is that the hard instances of this problem are those with small mean rewards, i.e., the small click-through rates that are most relevant in practice. Based on this, and the fact that small mean implies small variance for Bernoullis, our key technical result shows that variance-aware confidence sets derived from the Bernstein and Chernoff bounds lead to optimal algorithms (up to log terms), whereas Hoeffding-based algorithms suffer order-wise suboptimal regret. This sharply contrasts with the standard (non-cascading) bandit setting, where the variance-aware algorithms only improve constants. In light of this and as an additional contribution, we propose a variance-aware algorithm for the structured case of linear rewards and show its regret strictly improves the state-of-the-art.

## Sample-Efficient Reinforcement Learning of Partially Observable Markov Games

- Qinghua Liu · Csaba Szepesvari · Chi Jin
- abstract@[open-review](#): This paper considers the challenging tasks of Multi-Agent Reinforcement Learning (MARL) under partial observability, where each agent only sees her own individual observations and actions that reveal incomplete information about the underlying state of system. This paper studies these tasks under the general model of multiplayer general-sum Partially Observable Markov Games (POMGs), which is significantly larger than the standard model of Imperfect Information Extensive-Form Games (IIEFGs). We identify a rich subclass of POMGs---weakly revealing POMGs---in which sample-efficient learning is tractable. In the self-play setting, we prove that a simple algorithm combining optimism and Maximum Likelihood Estimation (MLE) is sufficient to find approximate Nash equilibria, correlated equilibria, as well as coarse correlated equilibria of weakly revealing POMGs, in a polynomial number of samples when the number of agents is small. In the setting of playing against adversarial opponents, we show that a variant of our optimistic MLE algorithm is capable of achieving sublinear regret when being compared against the optimal maximin policies. To our best knowledge, this work provides the first line of sample-efficient results for learning POMGs.

## Syndicated Bandits: A Framework for Auto Tuning Hyper-parameters in Contextual Bandit Algorithms

- QIN DING · Yue Kang · Yi-Wei Liu · Thomas Chun Man Lee · Cho-Jui Hsieh · James Sharpnack
- abstract@[open-review](#): The stochastic contextual bandit problem, which models the trade-off between exploration and exploitation, has many real applications, including recommender systems, online advertising and clinical trials. As many other machine learning algorithms, contextual bandit algorithms often have one or more hyper-parameters. As an example, in most optimal stochastic contextual bandit algorithms, there is an unknown exploration parameter which controls the trade-off between exploration and exploitation. A proper choice of the hyper-parameters is essential for contextual bandit algorithms to perform well. However, it is infeasible to use offline tuning methods to select hyper-parameters in contextual bandit environment since there is no pre-collected dataset and the decisions have to be made in real time. To tackle this problem, we first propose a two-layer bandit structure for auto tuning the exploration parameter and further generalize it to the Syndicated Bandits framework which can learn multiple hyper-parameters dynamically in contextual bandit environment. We derive the regret bounds of our proposed Syndicated Bandits framework and show it can avoid its regret dependent exponentially in the number of hyper-parameters to be tuned. Moreover, it achieves optimal regret bounds under certain

scenarios. Syndicated Bandits framework is general enough to handle the tuning tasks in many popular contextual bandit algorithms, such as LinUCB, LinTS, UCB-GLM, etc. Experiments on both synthetic and real datasets validate the effectiveness of our proposed framework.

## [Convergence for score-based generative modeling with polynomial complexity](#)

- Holden Lee · Jianfeng Lu · Yixin Tan
- abstract@[open-review](#): Score-based generative modeling (SGM) is a highly successful approach for learning a probability distribution from data and generating further samples. We prove the first polynomial convergence guarantees for the core mechanic behind SGM: drawing samples from a probability density  $p$  given a score estimate (an estimate of  $\nabla \ln p$ ) that is accurate in  $L^2(p)$ . Compared to previous works, we do not incur error that grows exponentially in time or that suffers from a curse of dimensionality. Our guarantee works for any smooth distribution and depends polynomially on its log-Sobolev constant. Using our guarantee, we give a theoretical analysis of score-based generative modeling, which transforms white-noise input into samples from a learned data distribution given score estimates at different noise scales. Our analysis gives theoretical grounding to the observation that an annealed procedure is required in practice to generate good samples, as our proof depends essentially on using annealing to obtain a warm start at each step. Moreover, we show that a predictor-corrector algorithm gives better convergence than using either portion alone.

## [The Neural Covariance SDE: Shaped Infinite Depth-and-Width Networks at Initialization](#)

- Mufan Li · Mihai Nica · Daniel M Roy
- abstract@[open-review](#): The logit outputs of a feedforward neural network at initialization are conditionally Gaussian, given a random covariance matrix defined by the penultimate layer. In this work, we study the distribution of this random matrix. Recent work has shown that shaping the activation function as network depth grows large is necessary for this covariance matrix to be non-degenerate. However, the current infinite-width-style understanding of this shaping method is unsatisfactory for large depth: infinite-width analyses ignore the microscopic fluctuations from layer to layer, but these fluctuations accumulate over many layers. To overcome this shortcoming, we study the random covariance matrix in the shaped infinite-depth-and-width limit. We identify the precise scaling of the activation function necessary to arrive at a non-trivial limit, and show that the random covariance matrix is governed by a stochastic differential equation (SDE) that we call the Neural Covariance SDE. Using simulations, we show that the SDE closely matches the distribution of the random covariance matrix of finite networks. Additionally, we recover an if-and-only-if condition for exploding and vanishing norms of large shaped networks based on the activation function.

## [Sketch-GNN: Scalable Graph Neural Networks with Sublinear Training Complexity](#)

- Mucong Ding · Tahseen Rabbani · Bang An · Evan Wang · Furong Huang
- abstract@[open-review](#): Graph Neural Networks (GNNs) are widely applied to graph learning problems such as node classification. When scaling up the underlying graphs of GNNs to a larger size, we are forced to either train on the complete graph and keep the full graph adjacency and node embeddings in memory (which is often infeasible) or mini-batch sample the graph (which results in exponentially growing computational complexities with respect to the number of GNN layers). Various sampling-based and historical-embedding-based methods are proposed to avoid this exponential growth of complexities. However, none of these solutions eliminates the linear dependence on graph size. This paper proposes a sketch-based algorithm whose training time and memory grow sublinearly with respect to graph size by training GNNs atop a few compact sketches of graph adjacency and node embeddings. Based on polynomial tensor-sketch (PTS) theory, our framework provides a novel protocol for sketching non-linear activations and graph convolution matrices in GNNs, as opposed to existing methods that sketch linear weights or gradients in neural networks. In addition, we develop a locality sensitive hashing (LSH) technique that can be trained to improve the quality of sketches. Experiments on large-graph benchmarks demonstrate the scalability and competitive performance of our Sketch-GNNs versus their full-size GNN counterparts.

## [Understanding Hyperdimensional Computing for Parallel Single-Pass Learning](#)

- Tao Yu · Yichi Zhang · Zhiru Zhang · Christopher De Sa
- abstract@[open-review](#): Hyperdimensional computing (HDC) is an emerging learning paradigm that computes with high dimensional binary vectors. There is an active line of research on HDC in the community of emerging hardware because of its energy efficiency and ultra-low latency---but HDC suffers from low model accuracy, with little theoretical understanding of what limits its performance. We propose a new theoretical analysis of the limits of HDC via a consideration of what similarity matrices can be expressed" by binary vectors, and we show how the limits of HDC can be approached using random Fourier features (RFF). We extend our analysis to the more general class of vector symbolic architectures (VSA), which compute with high-dimensional vectors (hypervectors) that are not necessarily binary. We propose a new class of VSAs, finite group VSAs, which surpass the limits of HDC. Using representation theory, we characterize which similarity matrices can be expressed" by finite group VSA hypervectors, and we show how these VSAs can be constructed. Experimental results show that our RFF method and group VSA can both outperform the state-of-the-art HDC model by up to 7.6% while maintaining hardware efficiency. This work aims to inspire a future interest on HDC in the ML community and connect to the hardware community.

## [Assessing representation quality in Self-Supervised Learning by measuring eigenspectrum decay](#)

- Kumar K Agrawal · Arnab Mondal · Arna Ghosh · Blake Richards
- abstract@[open-review](#): Self-supervised learning (SSL) with large-scale unlabelled datasets enables the learning of useful representations for multiple downstream tasks. However, assessing the quality of these representations is a challenging open problem. Directly evaluating performance on specific tasks is either inefficient (if measuring performance on many different tasks) or insufficient (if measuring performance on only a small number of tasks). This leaves us with a dilemma, how do we determine if representations learned with SSL will generalize well across a wide range of potential downstream tasks? Therefore, some task-agnostic statistical measure of representation quality for estimating generalization without explicit downstream task assessment would be highly desirable. Here, we analyze representations learned by deep neural networks, pretrained with multiple SSL learning objectives with the goal of developing a task-agnostic metric of representation quality. In-line with recent work, we find that the eigenspectrum of the empirical representation covariance matrix can be approximated with a power-law distribution. Furthermore, if a task is solvable with a linear readout from the representations, we find that the decay coefficient of the eigenspectrum ( $\alpha$ ) is a good measure of representation quality, i.e. there exists an interval for  $\alpha$  where representations exhibit excellent downstream task generalization. We provide analytical and empirical evidence across several datasets to support this claim. Notably,  $\alpha$  depends only on the representations themselves, and as such, it offers an efficient, general mechanism for assessing representation quality and guiding model selection in SSL.

## [FedRolex: Model-Heterogeneous Federated Learning with Rolling Submodel Extraction](#)

- Samiul Alam · Luyang Liu · Ming Yan · Mi Zhang
- abstract@[open-review](#): Federated learning (FL) is a collaborative machine learning paradigm to train models from decentralized private data. Most FL research focuses on the model-homogeneous setting where models deployed across all the participating clients and server is required to be identical. However, in real-world scenarios, such a requirement acts as a constraint that restricts the outreach to clients with heterogeneous device resources and unfairly excludes users with low-end devices who would otherwise benefit from FL. In this work, we propose a simple yet effective model-heterogeneous FL method named FedRolex to tackle this constraint. Unlike the model-homogeneous scenario, the fundamental challenge of model heterogeneity in FL is that different parameters of the global model are trained on heterogeneous data distributions. FedRolex addresses this challenge by rolling the submodel in each federated iteration so that the parameters of the global model are evenly trained on the global data distribution across all devices, making it more akin to model-homogeneous training. Our experiments show that FedRolex outperforms other state-of-the-art model-heterogeneous FL methods, especially

under high data-heterogeneity scenarios. We have conducted ablation studies to show that submodel rolling is an effective technique to reduce the gap between model-heterogeneous and the standard model-homogeneous settings. Lastly, we consider the distribution of client capabilities that is similar to real-world income distribution instead of the uniform distribution used in existing works. Our results show a consistent improvement in accuracies on low-end devices which enhances the inclusiveness of FL.

## [Memory Efficient Continual Learning with Transformers](#)

- Beyza Ermis · Giovanni Zappella · Martin Wistuba · Aditya Rawal · Cedric Archambeau
- abstract@[open-review](#): In many real-world scenarios, data to train machine learning models becomes available over time. Unfortunately, these models struggle to continually learn new concepts without forgetting what has been learnt in the past. This phenomenon is known as catastrophic forgetting and it is difficult to prevent due to practical constraints. For instance, the amount of data that can be stored or the computational resources that can be used might be limited. Moreover, applications increasingly rely on large pre-trained neural networks, such as pre-trained Transformers, since compute or data might not be available in sufficiently large quantities to practitioners to train from scratch. In this paper, we devise a method to incrementally train a model on a sequence of tasks using pre-trained Transformers and extending them with Adapters. Different than the existing approaches, our method is able to scale to a large number of tasks without significant overhead and allows sharing information across tasks. On both image and text classification tasks, we empirically demonstrate that our method maintains a good predictive performance without retraining the model or increasing the number of model parameters over time. The resulting model is also significantly faster at inference time compared to Adapter-based state-of-the-art methods.

## [Audio-Driven Co-Speech Gesture Image Generation](#)

- Xian Liu · Qianyi Wu · Hang Zhou · Yuanqi Du · Wayne Wu · Dahua Lin · Ziwei Liu
- abstract@[open-review](#): Co-speech gesture is crucial for human-machine interaction and digital entertainment. While previous works mostly map speech audio to human skeletons (e.g., 2D keypoints), directly generating speakers' gestures in the image domain remains unsolved. In this work, we formally define and study this challenging problem of audio-driven co-speech gesture image generation, i.e., using a unified framework to generate speaker image sequence driven by speech audio. Our key insight is that the co-speech gestures can be decomposed into common motion patterns and subtle rhythmic dynamics. To this end, we propose a novel framework, Audio-driveN Gesture Image gEneration (ANGIE), to effectively capture the reusable co-speech gesture patterns as well as fine-grained rhythmic movements. To achieve high-fidelity image sequence generation, we leverage an unsupervised motion representation instead of a structural human body prior (e.g., 2D skeletons). Specifically, 1) we propose a vector quantized motion extractor (VQ-Motion Extractor) to summarize common co-speech gesture patterns from implicit motion representation to codebooks. 2) Moreover, a co-speech gesture GPT with motion refinement (Co-Speech GPT) is devised to complement the subtle prosodic motion details. Extensive experiments demonstrate that our framework renders realistic and vivid co-speech gesture images. All the code, data and models will be released.

## [A Multi-Resolution Framework for U-Nets with Applications to Hierarchical VAEs](#)

- Fabian Falck · Christopher Williams · Dominic Danks · George Deligiannidis · Christopher Yau · Chris C Holmes · Arnaud Doucet · Matthew Willetts
- abstract@[open-review](#): U-Net architectures are ubiquitous in state-of-the-art deep learning, however their regularisation properties are understudied. In this paper, we formulate a multi-resolution framework which identifies U-Nets as finite-dimensional truncations of models on an infinite-dimensional function space. We provide theoretical results which prove that average pooling corresponds to projection within the space of square-integrable functions and show that U-Nets with average pooling implicitly learn a Haar wavelet basis representation of the data. We then leverage our framework to identify state-of-the-art hierarchical VAEs (HVAEs), which have a U-Net architecture, as forward Euler discretisations of multi-resolution diffusion sum processes which flow to a point mass, introducing instabilities. We also demonstrate that HVAEs learn a representation of time which allows for weight-sharing without detriment. We use this observation to achieve state-of-the-art HVAE performance with half the number of parameters of existing models, exploiting the properties of our continuous-time formulation.

## [Self-Supervised Pretraining for 3D Vision Tasks by Cross-View Completion](#)

- Philippe Weinzaepfel · Vincent Leroy · Thomas Lucas · Romain BRÃ‰GIER · Yohann Cabon · Vaibhav ARORA · Leonid Antsfeld · Boris Chidlovskii · Gabriela Csurka · Jerome Revaud
- abstract@[open-review](#): Masked Image Modeling (MIM) has recently been established as a potent pretraining paradigm. A pretext task is constructed by masking patches in an input image, and this masked content is then predicted by a neural network using visible patches as sole input. This pretraining leads to state-of-the-art performance when finetuned for high-level semantic tasks, e.g. image classification and object detection. In this paper we instead seek to learn representations that transfer well to a wide variety of 3D vision and lower-level geometric downstream tasks, such as depth prediction or optical flow estimation. Inspired by MIM, we propose an unsupervised representation learning task trained from pairs of images showing the same scene from different viewpoints. More precisely, we propose the pretext task of cross-view completion where the first input image is partially masked, and this masked content has to be reconstructed from the visible content and the second image. In single-view MIM the masked content often cannot be inferred precisely from the visible portion only, so the model learns to act as a prior influenced by high-level semantics. In contrast, this ambiguity can be resolved with cross-view completion from the second unmasked image, on the condition that the model is able to understand the spatial relationship between the two images. Our experiments show that our pretext task leads to significantly improved performance for monocular 3D vision downstream tasks such as depth estimation. In addition, our model can be, by its design, directly applied to binocular downstream tasks such as optical flow or relative camera pose estimation, for which we obtain competitive results without bells and whistles, i.e. using a generic architecture without any task-specific design.

## [Watermarking for Out-of-distribution Detection](#)

- Qizhou Wang · Feng Liu · Yonggang Zhang · Jing Zhang · Chen Gong · Tongliang Liu · Bo Han
- abstract@[open-review](#): Out-of-distribution (OOD) detection aims to identify OOD data based on representations extracted from well-trained deep models. However, existing methods largely ignore the reprogramming property of deep models and thus may not fully unleash their intrinsic strength: without modifying parameters of a well-trained deep model, we can reprogram this model for a new purpose via data-level manipulation (e.g., adding a specific feature perturbation). This property motivates us to reprogram a classification model to excel at OOD detection (a new task), and thus we propose a general methodology named watermarking in this paper. Specifically, we learn a unified pattern that is superimposed onto features of original data, and the model's detection capability is largely boosted after watermarking. Extensive experiments verify the effectiveness of watermarking, demonstrating the significance of the reprogramming property of deep models in OOD detection.

## [DOPE: Doubly Optimistic and Pessimistic Exploration for Safe Reinforcement Learning](#)

- Archana Bura · Aria HasanzadeZonuzy · Dileep Kalathil · Srinivas Shakkottai · Jean-Francois Chamberland
- abstract@[open-review](#): Safe reinforcement learning is extremely challenging. Not only must the agent explore an unknown environment, it must do so while ensuring no safety constraint violations. The problem is typically posed as a constrained Markov decision process (MDP) under an unknown model, often with the learning agent having access to a safe suboptimal baseline policy. Recent results obtain an  $\tilde{O}(\sqrt{S}IK)$  objective regret in  $K$  episodes for an MDP with  $S$  states, while being safe at all times. The main idea is to combine a reward bonus for exploration (optimism) with a conservative constraint (pessimism). However, the approach is so pessimistic that empirical results indicate an inordinately long learning process that keeps on applying sequences of the safe baseline policy. Our key insight is that such excessive pessimism hinders exploration,

and needs to be combated by optimism with respect to the model. This insight yields DOPE, which has a double dose of optimism with respect to model and reward, while being pessimistic with respect to the constraints. We show that DOPE reduces the objective regret to  $\tilde{O}(\sqrt{SK})$ , with no constraint violation. Furthermore, we show in empirical studies that DOPE has a dramatic performance improvement as compared to earlier approaches.

## [A Few Expert Queries Suffices for Sample-Efficient RL with Resets and Linear Value Approximation](#)

- Philip Amortila · Nan Jiang · Dean Foster · Dhruv Madenka
- abstract@[open-review](#): The current paper studies sample-efficient Reinforcement Learning (RL) in settings where only the optimal value function is assumed to be linearly-realizable. It has recently been understood that, even under this seemingly strong assumption and access to a generative model, worst-case sample complexities can be prohibitively (i.e., exponentially) large. We investigate the setting where the learner additionally has access to interactive demonstrations from an expert policy, and we present a statistically and computationally efficient algorithm (Delphi) for blending exploration with expert queries. In particular, Delphi requires  $\tilde{O}(d)$  expert queries and a  $\text{poly}(d, H, |A|, 1/\epsilon)$  amount of exploratory samples to provably recover an  $\epsilon$ -suboptimal policy. Compared to pure RL approaches, this corresponds to an exponential improvement in sample complexity with surprisingly-little expert input. Compared to prior imitation learning (IL) approaches, our required number of expert demonstrations is independent of  $H$  and logarithmic in  $1/\epsilon$ , whereas all prior work required at least linear factors of both in addition to the same dependence on  $d$ . Towards establishing the minimal amount of expert queries needed, we show that, in the same setting, any learner whose exploration budget is polynomially-bounded (in terms of  $d, H, |A|$ ) will require at least  $\tilde{\Omega}(\sqrt{d})$  oracle calls to recover a policy competing with the expert's value function. Under the weaker assumption that the expert's policy is linear, we show that the lower bound increases to  $\tilde{\Omega}(d)$ .

## [DaDA: Distortion-aware Domain Adaptation for Unsupervised Semantic Segmentation](#)

- Sujin Jang · Joohan Na · Dokwan Oh
- abstract@[open-review](#): Distributional shifts in photometry and texture have been extensively studied for unsupervised domain adaptation, but their counterparts in optical distortion have been largely neglected. In this work, we tackle the task of unsupervised domain adaptation for semantic image segmentation where unknown optical distortion exists between source and target images. To this end, we propose a distortion-aware domain adaptation (DaDA) framework that boosts the unsupervised segmentation performance. We first present a relative distortion learning (RDL) approach that is capable of modeling domain shifts in fine-grained geometric deformation based on diffeomorphic transformation. Then, we demonstrate that applying additional global affine transformations to the diffeomorphically transformed source images can further improve the segmentation adaptation. Besides, we find that our distortion-aware adaptation method helps to enhance self-supervised learning by providing higher-quality initial models and pseudo labels. To evaluate, we propose new distortion adaptation benchmarks, where rectilinear source images and fisheye target images are used for unsupervised domain adaptation. Extensive experimental results highlight the effectiveness of our approach over the state-of-the-art methods under unknown relative distortion across domains.

## [Deep Active Learning by Leveraging Training Dynamics](#)

- Haonan Wang · Wei Huang · Ziwei Wu · Hanghang Tong · Andrew J Margenot · Jingrui He
- abstract@[open-review](#): Active learning theories and methods have been extensively studied in classical statistical learning settings. However, deep active learning, i.e., active learning with deep learning models, is usually based on empirical criteria without solid theoretical justification, thus suffering from heavy doubts when some of those fail to provide benefits in applications. In this paper, by exploring the connection between the generalization performance and the training dynamics, we propose a theory-driven deep active learning method (dynamicAL) which selects samples to maximize training dynamics. In particular, we prove that the convergence speed of training and the generalization performance is positively correlated under the ultra-wide condition and show that maximizing the training dynamics leads to a better generalization performance. Furthermore, to scale up to large deep neural networks and data sets, we introduce two relaxations for the subset selection problem and reduce the time complexity from polynomial to constant. Empirical results show that dynamicAL not only outperforms the other baselines consistently but also scales well on large deep learning models. We hope our work inspires more attempts in bridging the theoretical findings of deep networks and practical impacts in deep active learning applications.

## [Adaptive Stochastic Variance Reduction for Non-convex Finite-Sum Minimization](#)

- Ali Kavousi · EFSTRATIOS SKOULAKIS · Kimon Antonakopoulos · Leello Tadesse Dadi · Volkan Cevher
- abstract@[open-review](#): We propose an adaptive variance-reduction method, called AdaSpider, for minimization of  $L$ -smooth, non-convex functions with a finite-sum structure. In essence, AdaSpider combines an AdaGrad-inspired (Duchi et al., 2011), but a fairly distinct, adaptive step-size schedule with the recursive stochastic path integrated estimator proposed in (Fang et al., 2018). To our knowledge, AdaSpider is the first parameter-free non-convex variance-reduction method in the sense that it does not require the knowledge of problem-dependent parameters, such as smoothness constant  $L$ , target accuracy  $\epsilon$  or any bound on gradient norms. In doing so, we are able to compute an  $\epsilon$ -stationary point with  $\tilde{O}(\sqrt{n} \log(n))$  oracle-calls, which matches the respective lower bound up to logarithmic factors.

## [Conformal Off-Policy Prediction in Contextual Bandits](#)

- Muhammad Faaiz Taufiq · Jean-Francois Ton · Rob Cornish · Yee Whye Teh · Arnaud Doucet
- abstract@[open-review](#): Most off-policy evaluation methods for contextual bandits have focused on the expected outcome of a policy, which is estimated via methods that at best provide only asymptotic guarantees. However, in many applications, the expectation may not be the best measure of performance as it does not capture the variability of the outcome. In addition, particularly in safety-critical settings, stronger guarantees than asymptotic correctness may be required. To address these limitations, we consider a novel application of conformal prediction to contextual bandits. Given data collected under a behavioral policy, we propose conformal off-policy prediction (COPP), which can output reliable predictive intervals for the outcome under a new target policy. We provide theoretical finite-sample guarantees without making any additional assumptions beyond the standard contextual bandit setup, and empirically demonstrate the utility of COPP compared with existing methods on synthetic and real-world data.

## [What Makes A Good Code? A New Look At LSH From Random Fourier Features](#)

- Xiaoyun Li · Ping Li
- abstract@[open-review](#): The method of Random Fourier Feature (RFF) has been popular for large-scale learning, which generates non-linear random features of the data. It has also been used to construct Locality-Sensitive Hashing (LSH) codes via stochastic quantization for efficient information retrieval. In this paper, we revisit binary hashing from RFF, and study SignRFF, a simple strategy to extract RFF-based binary codes. Particularly, we ask: what makes a good LSH binary code? We answer the question by investigating a new measure called ranking efficiency, which provides a systematic and unified framework for comparing different LSH methods in practice. It suggests that the non-linear kernel based methods (e.g., SignRFF) should be preferred over the simple LSH in high similarity region. Experiments are conducted to show that SignRFF is consistently better than the previous RFF-based method, and also outperforms other data-dependent and deep learning based hashing methods with sufficient number of hash bits. Moreover, the proposed ranking efficiency aligns well with the empirical search performance.

## [Training Uncertainty-Aware Classifiers with Conformalized Deep Learning](#)

- Bat-Sheva Einbinder · Yaniv Romano · Matteo Sesa · Yanfei Zhou
- abstract@[open-review](#): Deep neural networks are powerful tools to detect hidden patterns in data and leverage them to make predictions, but they are not designed to understand uncertainty and estimate reliable probabilities. In particular, they tend to be overconfident. We begin to address this problem in the context of multi-class classification by developing a novel training algorithm producing models with more dependable uncertainty estimates, without sacrificing predictive power. The idea is to mitigate overconfidence by minimizing a loss function, inspired by advances in conformal inference, that quantifies model uncertainty by carefully leveraging hold-out data. Experiments with synthetic and real data demonstrate this method can lead to smaller conformal prediction sets with higher conditional coverage, after exact calibration with hold-out data, compared to state-of-the-art alternatives.

## [Semantic Probabilistic Layers for Neuro-Symbolic Learning](#)

- Kareem Ahmed · Stefano Teso · Kai-Wei Chang · Guy Van den Broeck · Antonio Vergari
- abstract@[open-review](#): We design a predictive layer for structured-output prediction (SOP) that can be plugged into any neural network guaranteeing its predictions are consistent with a set of predefined symbolic constraints. Our Semantic Probabilistic Layer (SPL) can model intricate correlations, and hard constraints, over a structured output space all while being amenable to end-to-end learning via maximum likelihood. SPLs combine exact probabilistic inference with logical reasoning in a clean and modular way, learning complex distributions and restricting their support to solutions of the constraint. As such, they can faithfully, and efficiently, model complex SOP tasks beyond the reach of alternative neuro-symbolic approaches. We empirically demonstrate that SPLs outperform these competitors in terms of accuracy on challenging SOP tasks such as hierarchical multi-label classification, pathfinding and preference learning, while retaining perfect constraint satisfaction.

## [SketchBoost: Fast Gradient Boosted Decision Tree for Multioutput Problems](#)

- Leonid Iosipoi · Anton Vakhrushev
- abstract@[open-review](#): Gradient Boosted Decision Tree (GBDT) is a widely-used machine learning algorithm that has been shown to achieve state-of-the-art results on many standard data science problems. We are interested in its application to multioutput problems when the output is highly multidimensional. Although there are highly effective GBDT implementations, their scalability to such problems is still unsatisfactory. In this paper, we propose novel methods aiming to accelerate the training process of GBDT in the multioutput scenario. The idea behind these methods lies in the approximate computation of a scoring function used to find the best split of decision trees. These methods are integrated into our easily customizable GPU implementation of GBDT in Python which we call SketchBoost. Our numerical study demonstrates that SketchBoost speeds up the training process of GBDT by up to over 40 times while achieving comparable or even better performance.

## [Towards Improving Faithfulness in Abstractive Summarization](#)

- Xiuying Chen · Mingzhe Li · Xin Gao · Xiangliang Zhang
- abstract@[open-review](#): Despite the success achieved in neural abstractive summarization based on pre-trained language models, one unresolved issue is that the generated summaries are not always faithful to the input document. There are two possible causes of the unfaithfulness problem: (1) the summarization model fails to understand or capture the gist of the input text, and (2) the model over-relies on the language model to generate fluent but inadequate words. In this work, we propose a Faithfulness Enhanced Summarization model (FES), which is designed for addressing these two problems and improving faithfulness in abstractive summarization. For the first problem, we propose to use question-answering (QA) to examine whether the encoder fully grasps the input document and can answer the questions on the key information in the input. The QA attention on the proper input words can also be used to stipulate how the decoder should attend to the source. For the second problem, we introduce a max-margin loss defined on the difference between the language and the summarization model, aiming to prevent the overconfidence of the language model. Extensive experiments on two benchmark summarization datasets, CNN/DM and XSum, demonstrate that our model significantly outperforms strong baselines. The evaluation of factual consistency also shows that our model generates more faithful summaries than baselines.

## [Explaining a Reinforcement Learning Agent via Prototyping](#)

- Ronilo Ragodos · Qihang Lin · Xun Zhou · Tong Wang
- abstract@[open-review](#): While deep reinforcement learning has proven to be successful in solving control tasks, the ``black-box'' nature of an agent has received increasing concerns. We propose a prototype-based post-hoc \texttt{policy explainer}, ProtoX, that explains a black-box agent by prototyping the agent's behaviors into scenarios, each represented by a prototypical state. When learning prototypes, ProtoX considers both visual similarity and scenario similarity. The latter is unique to the reinforcement learning context since it explains why the same action is taken in visually different states. To teach ProtoX about visual similarity, we pre-train an encoder using contrastive learning via self-supervised learning to recognize states as similar if they occur close together in time and receive the same action from the black-box agent. We then add an isometry layer to allow ProtoX to adapt scenario similarity to the downstream task. ProtoX is trained via imitation learning using behavior cloning, and thus requires no access to the environment or agent. In addition to explanation fidelity, we design different prototype shaping terms in the objective function to encourage better interpretability. We conduct various experiments to test ProtoX. Results show that ProtoX achieved high fidelity to the original black-box agent while providing meaningful and understandable explanations.

## [A theory of weight distribution-constrained learning](#)

- Weishun Zhong · Ben Sorscher · Daniel Lee · Haim Sompolinsky
- abstract@[open-review](#): A central question in computational neuroscience is how structure determines function in neural networks. Recent large-scale connectomic studies have started to provide a wealth of structural information such as the distribution of excitatory/inhibitory cell and synapse types as well as the distribution of synaptic weights in the brains of different species. The emerging high-quality large structural datasets raise the question of what general functional principles can be gleaned from them. Motivated by this question, we developed a statistical mechanical theory of learning in neural networks that incorporates structural information as constraints. We derived an analytical solution for the memory capacity of the perceptron, a basic feedforward model of supervised learning, with constraint on the distribution of its weights. Interestingly, the theory predicts that the reduction in capacity due to the constrained weight-distribution is related to the Wasserstein distance between the cumulative distribution function of the constrained weights and that of the standard normal distribution. To test the theoretical predictions, we use optimal transport theory and information geometry to develop an SGD-based algorithm to find weights that simultaneously learn the input-output task and satisfy the distribution constraint. We show that training in our algorithm can be interpreted as geodesic flows in the Wasserstein space of probability distributions. We further developed a statistical mechanical theory for teacher-student perceptron rule learning and ask for the best way for the student to incorporate prior knowledge of the rule (i.e., the teacher). Our theory shows that it is beneficial for the learner to adopt different prior weight distributions during learning, and shows that distribution-constrained learning outperforms unconstrained and sign-constrained learning. Our theory and algorithm provide novel strategies for incorporating prior knowledge about weights into learning, and reveal a powerful connection between structure and function in neural networks.

## [Unsupervised Learning under Latent Label Shift](#)

- Manley Roberts · Pranav Mani · Saurabh Garg · Zachary Lipton

- abstract@[open-review](#): What sorts of structure might enable a learner to discover classes from unlabeled data? Traditional approaches rely on feature-space similarity and heroic assumptions on the data. In this paper, we introduce unsupervised learning under Latent Label Shift (LLS), where the label marginals  $p_d(y)$  shift but the class conditionals  $p(x|y)$  do not. This work instantiates a new principle for identifying classes: elements that shift together group together. For finite input spaces, we establish an isomorphism between LLS and topic modeling: inputs correspond to words, domains to documents, and labels to topics. Addressing continuous data, we prove that when each label's support contains a separable region, analogous to an anchor word, oracle access to  $p(d|x)$  suffices to identify  $p_d(y)$  and  $p_d(y|x)$  up to permutation. Thus motivated, we introduce a practical algorithm that leverages domain-discriminative models as follows: (i) push examples through domain discriminator  $p(d|x)$ ; (ii) discretize the data by clustering examples in  $p(d|x)$  space; (iii) perform non-negative matrix factorization on the discrete data; (iv) combine the recovered  $p(y|d)$  with the discriminator outputs  $p(d|x)$  to compute  $p_d(y|x)$ . With semisynthetic experiments, we show that our algorithm can leverage domain information to improve state of the art unsupervised classification methods. We reveal a failure mode of standard unsupervised classification methods when data-space similarity does not indicate true groupings, and show empirically that our method better handles this case. Our results establish a deep connection between distribution shift and topic modeling, opening promising lines for future work.

## [On Privacy and Personalization in Cross-Silo Federated Learning](#)

- Ken Liu · Shengyuan Hu · Steven Wu · Virginia Smith
- abstract@[open-review](#): Although differential privacy (DP) is a well-studied topic in cross-device federated learning (FL), there is a lack of work considering DP for cross-silo FL, a setting characterized by a limited number of clients each containing many data subjects. In cross-silo FL, usual notions of client-level privacy are less suitable as real-world privacy regulations typically concern in-silo data subjects rather than the silos themselves. In this work, we instead consider the more realistic notion of silo-specific item-level privacy, where silos set their own privacy targets for local examples. Under this setting we reconsider the roles of privacy and personalization in federated learning. In particular, we show that mean-regularized multi-task learning (MR-MTL), a simple personalization framework, is a surprisingly strong baseline for cross-silo FL: under stronger privacy, silos are further incentivized to ``federate'' with each other to mitigate DP noise, resulting in consistent improvements relative to standard cross-device baselines. We provide a thorough empirical study of competing methods as well as a theoretical characterization of MR-MTL for a mean estimation problem, highlighting the interplay between privacy and cross-silo data heterogeneity. Our work serves to establish baselines for private cross-silo FL as well as identify key directions for future work in this area.

## [Weisfeiler and Leman Go Walking: Random Walk Kernels Revisited](#)

- Nils M. Kriege
- abstract@[open-review](#): Random walk kernels have been introduced in seminal work on graph learning and were later largely superseded by kernels based on the Weisfeiler-Leman test for graph isomorphism. We give a unified view on both classes of graph kernels. We study walk-based node refinement methods and formally relate them to several widely-used techniques, including Morgan's algorithm for molecule canonization and the Weisfeiler-Leman test. We define corresponding walk-based kernels on nodes that allow fine-grained parameterized neighborhood comparison, reach Weisfeiler-Leman expressiveness, and are computed using the kernel trick. From this we show that classical random walk kernels with only minor modifications regarding definition and computation are as expressive as the widely-used Weisfeiler-Leman subtree kernel but support non-strict neighborhood comparison. We verify experimentally that walk-based kernels reach or even surpass the accuracy of Weisfeiler-Leman kernels in real-world classification tasks.

## [FiLM-Ensemble: Probabilistic Deep Learning via Feature-wise Linear Modulation](#)

- Mehmet Ozgur Turkoglu · Alexander Becker · Huseyin Anil Gündüz · Mina Rezaei · Bernd Bischl · Rodrigo Caye Daudt · Stefano D'Aronco · Jan D. Wegner · Konrad Schindler
- abstract@[open-review](#): The ability to estimate epistemic uncertainty is often crucial when deploying machine learning in the real world, but modern methods often produce overconfident, uncalibrated uncertainty predictions. A common approach to quantify epistemic uncertainty, usable across a wide class of prediction models, is to train a model ensemble. In a naive implementation, the ensemble approach has high computational cost and high memory demand. This challenges in particular modern deep learning, where even a single deep network is already demanding in terms of compute and memory, and has given rise to a number of attempts to emulate the model ensemble without actually instantiating separate ensemble members. We introduce FiLM-Ensemble, a deep, implicit ensemble method based on the concept of Feature-wise Linear Modulation (FiLM). That technique was originally developed for multi-task learning, with the aim of decoupling different tasks. We show that the idea can be extended to uncertainty quantification: by modulating the network activations of a single deep network with FiLM, one obtains a model ensemble with high diversity, and consequently well-calibrated estimates of epistemic uncertainty, with low computational overhead in comparison. Empirically, FiLM-Ensemble outperforms other implicit ensemble methods, and it comes very close to the upper bound of an explicit ensemble of networks (sometimes even beating it), at a fraction of the memory cost.

## [Adversarial Auto-Augment with Label Preservation: A Representation Learning Principle Guided Approach](#)

- Kaiwen Yang · Yanchao Sun · Jiahao Su · Fengxiang He · Xinmei Tian · Furong Huang · Tianyi Zhou · Dacheng Tao
- abstract@[open-review](#): Data augmentation is a critical contributing factor to the success of deep learning but heavily relies on prior domain knowledge which is not always available. Recent works on automatic data augmentation learn a policy to form a sequence of augmentation operations, which are still pre-defined and restricted to limited options. In this paper, we show that a prior-free autonomous data augmentation's objective can be derived from a representation learning principle that aims to preserve the minimum sufficient information of the labels. Given an example, the objective aims at creating a distant ``hard positive example'' as the augmentation, while still preserving the original label. We then propose a practical surrogate to the objective that can be optimized efficiently and integrated seamlessly into existing methods for a broad class of machine learning tasks, e.g., supervised, semi-supervised, and noisy-label learning. Unlike previous works, our method does not require training an extra generative model but instead leverages the intermediate layer representations of the end-task model for generating data augmentations. In experiments, we show that our method consistently brings non-trivial improvements to the three aforementioned learning tasks from both efficiency and final performance, either or not combined with pre-defined augmentations, e.g., on medical images when domain knowledge is unavailable and the existing augmentation techniques perform poorly. Code will be released publicly.

## [Towards Disentangling Information Paths with Coded ResNeXt](#)

- Apostolos Avranas · Marios Kountouris
- abstract@[open-review](#): The conventional, widely used treatment of deep learning models as black boxes provides limited or no insights into the mechanisms that guide the neural network decisions. Significant research effort has been dedicated to building interpretable models to address this issue. Most efforts either focus on the high-level features associated with the last layers, or attempt to interpret the output of a single layer. In this paper, we take a novel approach to enhance the transparency of the function of the whole network. We propose a neural network architecture for classification in which the information that is relevant to each class flows through specific paths. These paths are designed in advance before training leveraging coding theory. Moreover, the paths do not depend on the semantic similarities between classes. A key property is that each path can be used as an autonomous single-purpose model. This enables to obtain, without any additional training and for any class, a lightweight binary classifier that has at least 60% fewer parameters than the original network. Furthermore, our coding theory based approach allows the neural network to make early predictions at intermediate layers during inference, without requiring its full evaluation. Remarkably, the proposed architecture provides all the aforementioned properties while significantly improving the overall accuracy. We demonstrate these properties on a slightly modified ResNeXt model tested on CIFAR-10/100 and ImageNet-1k.

## [How Mask Matters: Towards Theoretical Understandings of Masked Autoencoders](#)

- Qi Zhang · Yifei Wang · Yisen Wang
- abstract@[open-review](#): Masked AutoEncoder (MAE) based on a reconstruction task has risen to be a promising paradigm for self-supervised learning and achieves state-of-the-art performance across different benchmark datasets. However, despite its impressive empirical success, a theoretical analysis of it is still limited. In this paper, we propose a new theoretical understanding of how MAE works and why the choice of mask ratio is so important for MAE from a graph perspective. Based on the analysis of the MAE loss, we prove that the MAE loss can be upper bounded by an implicit alignment loss and propose an insight that MAE bridges different samples in the same class with an aggressive mask ratio. Then, we establish a guarantee for the downstream performance of MAE and analyze the trade-off on choosing the mask ratio. Motivated by our theory, we propose a Uniformity-promoting MAE (U-MAE) loss and find it can significantly improve the downstream performance of MAE on real-world datasets, including CIFAR-10 and ImageNet-100.

## [Online Neural Sequence Detection with Hierarchical Dirichlet Point Process](#)

- Weihan Li · Yu Qi · Gang Pan
- abstract@[open-review](#): Neural sequence detection plays an important role in neuroscience research. Recent impressive works utilize convolutive nonnegative matrix factorization and Neyman-Scott process to solve this problems. However, they still face two limitations. Firstly, they accommodate the entire dataset into memory and perform iterative updates of multiple passes, which can be inefficient when dataset is large or grows frequently. Secondly, they rely on the prior knowledge of the number of sequence types, which can be impractical with real data when the future situation is unknown. To tackle these limitations, we propose a hierarchical Dirichlet point processes model for efficient neural sequence detection. Instead of computing the entire data, our model can sequentially detect sequences in an online unsupervised manner with Particle filter. Besides, the Dirichlet prior enables our model to automatically introduce new sequence types on the fly as needed, thus avoiding specifying the number of types in advance. We manifest these advantages on synthetic data and real-world recordings from songbird higher vocal center and rodent hippocampus.

## [Modeling Transitivity and Cyclicity in Directed Graphs via Binary Code Box Embeddings](#)

- Dongxu Zhang · Michael Boratko · Cameron Musco · Andrew McCallum
- abstract@[open-review](#): Modeling directed graphs in continuous space is a fundamental requirement for performing machine learning on graph-structured data. Geometric embedding models (e.g., hyperbolic, cone, and box embeddings) excel at this task, exhibiting useful inductive biases for directed graphs. However, modeling directed graphs that both contain cycles and have transitive structure, two properties common in real-world settings, is challenging. Box embeddings, which can be thought of as representing the graph as an intersection over some learned subgraphs, have a natural inductive bias toward modeling transitive edges, but (as we prove) cannot model cycles. To this end, we propose binary code box embeddings, where a learned binary code selects a subset of graphs for intersection. We explore several variants, including global binary codes (amounting to a union over intersections) and per-vertex binary codes (allowing greater flexibility) as well as methods of regularization. Theoretical and empirical results show that the proposed models not only preserve a useful inductive bias of transitivity but also have sufficient representational capacity to model arbitrary graphs, including graphs with cycles.

## [Structural Pruning via Latency-Saliency Knapsack](#)

- Maying Shen · Hongxu Yin · Pavlo Molchanov · Lei Mao · Jianna Liu · Jose M. Alvarez
- abstract@[open-review](#): Structural pruning can simplify network architecture and improve inference speed. We propose Latency-Aware Structural Pruning (LASP) that formulates structural pruning as a global resource allocation optimization problem, aiming at maximizing the accuracy while constraining latency under a predefined budget. For filter importance ranking, LASP leverages latency lookup table to track latency reduction potential and global saliency score to gauge accuracy drop. Both metrics can be evaluated very efficiently during pruning, allowing us to reformulate global structural pruning under a reward maximization problem given target constraint. This makes the problem solvable via our augmented knapsack solver, enabling LASP to surpass prior work in pruning efficacy and accuracy-efficiency trade-off. We examine LASP on both classification and detection tasks, over varying networks, on ImageNet and VOC datasets, on different platforms. In particular, for ResNet-50/-101 pruning on ImageNet, LASP improves network throughput by \$1.60\times\$/\$1.90\times\$ with \$+0.3\%\$/\$-0.2\%\$ top-1 accuracy changes, respectively. For SSD pruning on VOC, LASP improves throughput by \$1.94\times\$ with only a \$0.56\%\$ mAP drop. LASP consistently outperforms prior art, sometimes by large margins.

## [Precise Regret Bounds for Log-loss via a Truncated Bayesian Algorithm](#)

- Changlong Wu · Mohsen Heidari · Ananth Grama · Wojciech Szpankowski
- abstract@[open-review](#): We study sequential general online regression, known also as sequential probability assignments, under logarithmic loss when compared against a broad class of experts. We obtain tight, often matching, lower and upper bounds for sequential minimax regret, which is defined as the excess loss incurred by the predictor over the best expert in the class. After proving a general upper bound we consider some specific classes of experts from Lipschitz class to bounded Hessian class and derive matching lower and upper bounds with provably optimal constants. Our bounds work for a wide range of values of the data dimension and the number of rounds. To derive lower bounds, we use tools from information theory (e.g., Shtarkov sum) and for upper bounds, we resort to new "smooth truncated covering" of the class of experts. This allows us to find constructive proofs by applying a simple and novel truncated Bayesian algorithm. Our proofs are substantially simpler than the existing ones and yet provide tighter (and often optimal) bounds.

## [\\$k\\\$-Sliced Mutual Information: A Quantitative Study of Scalability with Dimension](#)

- Ziv Goldfeld · Kristjan Greenewald · Theshani Nuradha · Galen Reeves
- abstract@[open-review](#): Sliced mutual information (SMI) is defined as an average of mutual information (MI) terms between one-dimensional random projections of the random variables. It serves as a surrogate measure of dependence to classic MI that preserves many of its properties but is more scalable to high dimensions. However, a quantitative characterization of how SMI itself and estimation rates thereof depend on the ambient dimension, which is crucial to the understanding of scalability, remain obscure. This work extends the original SMI definition to \$k\\$-\text{SMI}\$, which considers projections to \$k\\$-\text{dimensional subspaces}\$, and provides a multifaceted account on its dependence on dimension. Using a new result on the continuity of differential entropy in the 2-Wasserstein metric, we derive sharp bounds on the error of Monte Carlo (MC)-based estimates of \$k\\$-\text{SMI}\$, with explicit dependence on \$k\$ and the ambient dimension, revealing their interplay with the number of samples. We then combine the MC integrator with the neural estimation framework to provide an end-to-end \$k\\$-\text{SMI}\$ estimator, for which optimal convergence rates are established. We also explore asymptotics of the population \$k\\$-\text{SMI}\$ as dimension grows, providing Gaussian approximation results with a residual that decays under appropriate moment bounds. Our theory is validated with numerical experiments and is applied to sliced InfoGAN, which altogether provide a comprehensive quantitative account of the scalability question of \$k\\$-\text{SMI}\$, including SMI as a special case when \$k=1\$.

## [On Infinite Separations Between Simple and Optimal Mechanisms](#)

- Alexandros Psomas · Ariel Schwartzman Cohen · S. Weinberg
- abstract@[open-review](#): We consider a revenue-maximizing seller with \$k\$ heterogeneous items for sale to a single additive buyer, whose values are drawn from a known, possibly correlated prior \$\mathcal{D}\$. It is known that there exist priors \$\mathcal{D}\$ such that simple mechanisms --- those with bounded menu complexity --- extract an arbitrarily small fraction of the optimal revenue~(Briest et al. 2015, Hart and Nisan 2019). This paper considers

the opposite direction: given a correlated distribution  $\mathcal{D}$  witnessing an infinite separation between simple and optimal mechanisms, what can be said about  $\mathcal{D}$ ? [cit{hart2019selling}](#) provides a framework for constructing such  $\mathcal{D}$ : it takes as input a sequence of  $k$ -dimensional vectors satisfying some geometric property, and produces a  $\mathcal{D}$  witnessing an infinite gap. Our first main result establishes that this framework is without loss: *every*  $\mathcal{D}$  witnessing an infinite separation could have resulted from this framework. An earlier version of their work provided a more streamlined framework (Hart and Nisan 2013). Our second main result establishes that this restrictive framework is *not* tight. That is, we provide an instance  $\mathcal{D}$  witnessing an infinite gap, but which provably could not have resulted from the restrictive framework. As a corollary, we discover a new kind of mechanism which can witness these infinite separations on instances where the previous “aligned” mechanisms do not.

## Understanding and Extending Subgraph GNNs by Rethinking Their Symmetries

- Fabrizio Frasca · Beatrice Bevilacqua · Michael Bronstein · Haggai Maron
- abstract@[open-review](#): Subgraph GNNs are a recent class of expressive Graph Neural Networks (GNNs) which model graphs as collections of subgraphs. So far, the design space of possible Subgraph GNN architectures as well as their basic theoretical properties are still largely unexplored. In this paper, we study the most prominent form of subgraph methods, which employs node-based subgraph selection policies such as ego-networks or node marking and deletion. We address two central questions: (1) What is the upper-bound of the expressive power of these methods? and (2) What is the family of equivariant message passing layers on these sets of subgraphs?. Our first step in answering these questions is a novel symmetry analysis which shows that modelling the symmetries of node-based subgraph collections requires a significantly smaller symmetry group than the one adopted in previous works. This analysis is then used to establish a link between Subgraph GNNs and Invariant Graph Networks (IGNs). We answer the questions above by first bounding the expressive power of subgraph methods by 3-WL, and then proposing a general family of message-passing layers for subgraph methods that generalises all previous node-based Subgraph GNNs. Finally, we design a novel Subgraph GNN dubbed SUN, which theoretically unifies previous architectures while providing better empirical performance on multiple benchmarks.

## Learning Modular Simulations for Homogeneous Systems

- Jayesh Gupta · Sai Vemprala · Ashish Kapoor
- abstract@[open-review](#): Complex systems are often decomposed into modular subsystems for engineering tractability. Although various equation based white-box modeling techniques make use of such structure, learning based methods have yet to incorporate these ideas broadly. We present a modular simulation framework for modeling homogeneous multibody dynamical systems, which combines ideas from graph neural networks and neural differential equations. We learn to model the individual dynamical subsystem as a neural ODE module. Full simulation of the composite system is orchestrated via spatio-temporal message passing between these modules. An arbitrary number of modules can be combined to simulate systems of a wide variety of coupling topologies. We evaluate our framework on a variety of systems and show that message passing allows coordination between multiple modules over time for accurate predictions and in certain cases, enables zero-shot generalization to new system configurations. Furthermore, we show that our models can be transferred to new system configurations with lower data requirement and training effort, compared to those trained from scratch.

## Operative dimensions in unconstrained connectivity of recurrent neural networks

- Renate Krause · Matthew Cook · Sepp Kollmorgen · Valerio Mante · Giacomo Indiveri
- abstract@[open-review](#): Recurrent Neural Networks (RNN) are commonly used models to study neural computation. However, a comprehensive understanding of how dynamics in RNN emerge from the underlying connectivity is largely lacking. Previous work derived such an understanding for RNN fulfilling very specific constraints on their connectivity, but it is unclear whether the resulting insights apply more generally. Here we study how network dynamics are related to network connectivity in RNN trained without any specific constraints on several tasks previously employed in neuroscience. Despite the apparent high-dimensional connectivity of these RNN, we show that a low-dimensional, functionally relevant subspace of the weight matrix can be found through the identification of *operative* dimensions, which we define as components of the connectivity whose removal has a large influence on local RNN dynamics. We find that a weight matrix built from only a few operative dimensions is sufficient for the RNN to operate with the original performance, implying that much of the high-dimensional structure of the trained connectivity is functionally irrelevant. The existence of a low-dimensional, operative subspace in the weight matrix simplifies the challenge of linking connectivity to network dynamics and suggests that independent network functions may be placed in specific, separate subspaces of the weight matrix to avoid catastrophic forgetting in continual learning.

## Biologically plausible solutions for spiking networks with efficient coding

- Veronika Koren · Stefano Panzeri
- abstract@[open-review](#): Understanding how dynamics of neural networks is shaped by the computations the networks perform is a fundamental topic in neuroscience. Recently, the framework of efficient coding proposed a theory on how spiking neural networks can compute low-dimensional stimulus signals with high efficiency. Efficient spiking networks are based on time-dependent minimization of a cost function related to information coding with spikes. To inform the understanding of the function and dynamics of biological networks in the brain, however, the mathematical models have to obey constraints of biological networks. Currently, spiking network models of efficient coding have been extended to include some features of biological plausibility, such as the inclusion of architectures with excitatory and inhibitory neurons. However, biological realism of efficient coding theories is still limited to simple cases and does not include single neuron and network properties that are known to be important in biological circuits. Here, we revisit the theory of efficient coding with spikes to develop spiking neural networks that are closer to biological circuits. Namely, we find a biologically plausible spiking model realizing efficient coding in the case of a generalized leaky integrate-and-fire network with excitatory and inhibitory units, equipped with fast and slow synaptic currents, local homeostatic currents such as spike-triggered adaptation, hyperpolarization-activated rebound current, heterogeneous firing thresholds and resets, heterogeneous postsynaptic potentials, and structured, low-rank connectivity. We show how the complexity of E-E connectivity matrix shapes network responses.

## A Theoretical View on Sparsely Activated Networks

- Cenk Baykal · Nishanth Dikkala · Rina Panigrahy · Cyrus Rashtchian · Xin Wang
- abstract@[open-review](#): Deep and wide neural networks successfully fit very complex functions today, but dense models are starting to be prohibitively expensive for inference. To mitigate this, one promising research direction is networks that activate a sparse subgraph of the network. The subgraph is chosen by a data-dependent routing function, enforcing a fixed mapping of inputs to subnetworks (e.g., the Mixture of Experts (MoE) paradigm in Switch Transformers). However, there is no theoretical grounding for these sparsely activated models. As our first contribution, we present a formal model of data-dependent sparse networks that captures salient aspects of popular architectures. Then, we show how to construct sparse networks that provably match the approximation power and total size of dense networks on Lipschitz functions. The sparse networks use much fewer inference operations than dense networks, leading to a faster forward pass. The key idea is to use locality sensitive hashing on the input vectors and then interpolate the function in subregions of the input space. This offers a theoretical insight into why sparse networks work well in practice. Finally, we present empirical findings that support our theory; compared to dense networks, sparse networks give a favorable trade-off between number of active units and approximation quality.

## A Stochastic Linearized Augmented Lagrangian Method for Decentralized Bilevel Optimization

- Songtao Lu · Siliang Zeng · Xiaodong Cui · Mark Squillante · Lior Horesh · Brian Kingsbury · Jia Liu · Mingyi Hong

- abstract@[open-review](#): Bilevel optimization has been shown to be a powerful framework for formulating multi-task machine learning problems, e.g., reinforcement learning (RL) and meta-learning, where the decision variables are coupled in both levels of the minimization problems. In practice, the learning tasks would be located at different computing resource environments, and thus there is a need for deploying a decentralized training framework to implement multi-agent and multi-task learning. We develop a stochastic linearized augmented Lagrangian method (SLAM) for solving general nonconvex bilevel optimization problems over a graph, where both upper and lower optimization variables are able to achieve a consensus. We also establish that the theoretical convergence rate of the proposed SLAM to the Karush-Kuhn-Tucker (KKT) points of this class of problems is on the same order as the one achieved by the classical distributed stochastic gradient descent for only single-level nonconvex minimization problems. Numerical results tested on multi-agent RL problems showcase the superiority of SLAM compared with the benchmarks.

## [Predicting Single-Cell Perturbation Responses for Unseen Drugs](#)

- Leon Hetzel · Simon Boehm · Niki Kilbertus · Stephan Gähnemann · mohammad lotfollahi · Fabian Theis
- abstract@[open-review](#): Single-cell transcriptomics enabled the study of cellular heterogeneity in response to perturbations at the resolution of individual cells. However, scaling high-throughput screens (HTSs) to measure cellular responses for many drugs remains a challenge due to technical limitations and, more importantly, the cost of such multiplexed experiments. Thus, transferring information from routinely performed bulk RNA HTS is required to enrich single-cell data meaningfully. We introduce a new encoder-decoder architecture to study the perturbational effects of unseen drugs. We combine the model with an architecture surgery for transfer learning and demonstrate how training on existing bulk RNA HTS datasets can improve generalisation performance. Better generalisation reduces the need for extensive and costly screens at single-cell resolution. We envision that our proposed method will facilitate more efficient experiment designs through its ability to generate in-silico hypotheses, ultimately accelerating drug discovery.

## [Score-Based Models Detect Manifolds](#)

- Jakiw Pidstrigach
- abstract@[open-review](#): Score-based generative models (SGMs) need to approximate the scores  $\nabla \log p_t$  of the intermediate distributions as well as the final distribution  $p_T$  of the forward process. The theoretical underpinnings of the effects of these approximations are still lacking. We find precise conditions under which SGMs are able to produce samples from an underlying (low-dimensional) data manifold  $\mathcal{M}$ . This assures us that SGMs are able to generate the "right kind of samples". For example, taking  $\mathcal{M}$  to be the subset of images of faces, we provide conditions under which the SGM robustly produces an image of a face, even though the relative frequencies of these images might not accurately represent the true data generating distribution. Moreover, this analysis is a first step towards understanding the generalization properties of SGMs: Taking  $\mathcal{M}$  to be the set of all training samples, our results provide a precise description of when the SGM memorizes its training data.

## [Sparsity in Continuous-Depth Neural Networks](#)

- Hananeh Aliee · Till Richter · Mikhail Solonin · Ignacio Ibarra · Fabian Theis · Niki Kilbertus
- abstract@[open-review](#): Neural Ordinary Differential Equations (NODEs) have proven successful in learning dynamical systems in terms of accurately recovering the observed trajectories. While different types of sparsity have been proposed to improve robustness, the generalization properties of NODEs for dynamical systems beyond the observed data are underexplored. We systematically study the influence of weight and feature sparsity on forecasting as well as on identifying the underlying dynamical laws. Besides assessing existing methods, we propose a regularization technique to sparsify ``input-output connections'' and extract relevant features during training. Moreover, we curate real-world datasets including human motion capture and human hematopoiesis single-cell RNA-seq data to realistically analyze different levels of out-of-distribution (OOD) generalization in forecasting and dynamics identification respectively. Our extensive empirical evaluation on these challenging benchmarks suggests that weight sparsity improves generalization in the presence of noise or irregular sampling. However, it does not prevent learning spurious feature dependencies in the inferred dynamics, rendering them impractical for predictions under interventions, or for inferring the true underlying dynamics. Instead, feature sparsity can indeed help with recovering sparse ground-truth dynamics compared to unregularized NODEs.

## [SAGDA: Achieving \$\mathcal{O}\(\epsilon^{-2}\)\$ Communication Complexity in Federated Min-Max Learning](#)

- Haibo Yang · Zhuqing Liu · Xin Zhang · Jia Liu
- abstract@[open-review](#): Federated min-max learning has received increasing attention in recent years thanks to its wide range of applications in various learning paradigms. Similar to the conventional federated learning for empirical risk minimization problems, communication complexity also emerges as one of the most critical concerns that affects the future prospect of federated min-max learning. To lower the communication complexity of federated min-max learning, a natural approach is to utilize the idea of infrequent communications (through multiple local updates) same as in conventional federated learning. However, due to the more complicated inter-outer problem structure in federated min-max learning, theoretical understandings of communication complexity for federated min-max learning with infrequent communications remain very limited in the literature. This is particularly true for settings with non-i.i.d. datasets and partial client participation. To address this challenge, in this paper, we propose a new algorithmic framework called stochastic sampling averaging gradient descent ascent (SAGDA), which i) assembles stochastic gradient estimators from randomly sampled clients as control variates and ii) leverages two learning rates on both server and client sides. We show that SAGDA achieves a linear speedup in terms of both the number of clients and local update steps, which yields an  $\mathcal{O}(\epsilon^{-2})$  communication complexity that is orders of magnitude lower than the state of the art. Interestingly, by noting that the standard federated stochastic gradient descent ascent (FSGDA) is in fact a control-variate-free special version of SAGDA, we immediately arrive at an  $\mathcal{O}(\epsilon^{-2})$  communication complexity result for FSGDA. Therefore, through the lens of SAGDA, we also advance the current understanding on communication complexity of the standard FSGDA method for federated min-max learning.

## [FP8 Quantization: The Power of the Exponent](#)

- Andrey Kuzmin · Mart van Baalen · Yuwei Ren · Markus Nagel · Jorn Peters · Tijmen Blankevoort
- abstract@[open-review](#): When quantizing neural networks for efficient inference, low-bit integers are the go-to format for efficiency. However, low-bit floating point numbers have an extra degree of freedom, assigning some bits to work on an exponential scale instead. This paper exhaustively investigates this benefit of the floating point format for neural network inference. We detail the choices that can be made for the FP8 format, including the important choice of the number of bits for the mantissa and exponent, and show analytically in which settings these choices give better performance. Then we show how these findings translate to real networks, provide an efficient implementation for FP8 simulation, and a new algorithm that enables the learning of both the scale parameters and the number of exponent bits in the FP8 format. Our chief conclusion is that when doing post-training quantization for a wide range of networks, the FP8 format is better than INT8 in terms of accuracy, and the choice of the number of exponent bits is driven by the severity of outliers in the network. We also conduct experiments with quantization-aware training where the difference in formats disappears as the network is trained to reduce the effect of outliers.

## [Near-Optimal Regret Bounds for Multi-batch Reinforcement Learning](#)

- Zihan Zhang · Yuhang Jiang · Yuan Zhou · Xiangyang Ji
- abstract@[open-review](#): In this paper, we study the episodic reinforcement learning (RL) problem modeled by finite-horizon Markov Decision Processes (MDPs) with constraint on the number of batches. The multi-batch reinforcement learning framework, where the agent is required to provide a time schedule to update policy before everything, which is particularly suitable for the scenarios where the agent suffers extensively from changing the policy

adaptively. Given a finite-horizon MDP with  $S$  states,  $A$  actions and planning horizon  $H$ , we achieve near-optimal regret of  $\tilde{O}(\sqrt{SAH^3K\ln(1/\delta)})$  which hides logarithmic terms of  $(S,A,H,K)$  in  $K$  episodes using  $O(H + \log_2 K)$  batches with confidence parameter  $\delta$ . To our best of knowledge, it is the first  $\tilde{O}(\mathcal{P}(S,A,H)\sqrt{K})$  regret bound with  $O(H + \log_2 K)$  batch complexity. Meanwhile, we show that to achieve  $\tilde{O}(\mathcal{P}(S,A,H)\sqrt{K})$  regret, the number of batches is at least  $\Omega(H\log_A(K) + \log_2 K)$ , which matches our upper bound up to logarithmic terms. Our technical contribution are two-fold: 1) a near-optimal design scheme to explore over the unlearned states; 2) an computational efficient algorithm to explore certain directions with an approximated transition model.

## [Out-of-Distribution Detection via Conditional Kernel Independence Model](#)

- Yu Wang · Jingjing Zou · Jingyang Lin · Qing Ling · Yingwei Pan · Ting Yao · Tao Mei
- abstract@[open-review](#): Recently, various methods have been introduced to address the OOD detection problem with training outlier exposure. These methods usually count on discriminative softmax metric or energy method to screen OOD samples. In this paper, we probe an alternative hypothesis on OOD detection by constructing a novel latent variable model based on independent component analysis (ICA) techniques. This novel method named Conditional-i builds upon the probabilistic formulation, and applies the Hilbert-Schmidt Independence Criteria that offers a convenient solution for optimizing variable dependencies. Conditional-i exclusively encodes the useful class condition into the probabilistic model, which provides the desired convenience in delivering theoretical support for the OOD detection task. To facilitate the implementation of the Conditional-i model, we construct unique memory bank architectures that allow for convenient end-to-end training within a tractable budget. Empirical results demonstrate an evident performance boost on benchmarks against SOTA methods. We also provide valuable theoretical justifications that our training strategy is guaranteed to bound the error in the context of OOD detection. Code is available at: <https://github.com/anonymousneurips/conditional-i>.

## [Neural Abstractions](#)

- Alessandro Abate · Alec Edwards · Mirco Giacobbe
- abstract@[open-review](#): We present a novel method for the safety verification of non-linear dynamical systems. We train a neural network so as to approximate the system from sample states whilst ensuring an arbitrarily tight bound on the approximation error, which we formally certify using symbolic reasoning. If the latter step refutes the bound, then we augment the samples set with a counterexample and repeat training in a counterexample-guided inductive synthesis loop (CEGIS). We show that, upon successful certification of the bound, this produces a neural ODE with bounded disturbances that constitutes a formal abstraction of the original system, which satisfies a fundamental property: if the abstract system is safe then the original system is safe. Neural networks have extensively been used before as approximators; in this work we make a step further and use them for the first time as abstractions. By using neural ODEs with ReLU activation functions as abstractions, we cast the verification problem for non-linear systems into that of hybrid automata with affine dynamics. We demonstrate that our overall approach is particularly effective for the verification of benchmarks that do not exhibit Lipschitz continuity, which are out of reach to many existing technologies. Moreover, we demonstrate that it performs comparably to the mature tool for non-linear systems Flow\* over Lipschitz continuous examples.

## [Information-Theoretic Safe Exploration with Gaussian Processes](#)

- Alessandro Bottero · Carlos Luis · Julia Vinogradska · Felix Berkenkamp · Jan Peters
- abstract@[open-review](#): We consider a sequential decision making task where we are not allowed to evaluate parameters that violate an a priori unknown (safety) constraint. A common approach is to place a Gaussian process prior on the unknown constraint and allow evaluations only in regions that are safe with high probability. Most current methods rely on a discretization of the domain and cannot be directly extended to the continuous case. Moreover, the way in which they exploit regularity assumptions about the constraint introduces an additional critical hyperparameter. In this paper, we propose an information-theoretic safe exploration criterion that directly exploits the GP posterior to identify the most informative safe parameters to evaluate. Our approach is naturally applicable to continuous domains and does not require additional hyperparameters. We theoretically analyze the method and show that we do not violate the safety constraint with high probability and that we explore by learning about the constraint up to arbitrary precision. Empirical evaluations demonstrate improved data-efficiency and scalability.

## [Provably Adversarially Robust Detection of Out-of-Distribution Data \(Almost\) for Free](#)

- Alexander Meinke · Julian Bitterwolf · Matthias Hein
- abstract@[open-review](#): The application of machine learning in safety-critical systems requires a reliable assessment of uncertainty. However, deep neural networks are known to produce highly overconfident predictions on out-of-distribution (OOD) data. Even if trained to be non-confident on OOD data one can still adversarially manipulate OOD data so that the classifier again assigns high confidence to the manipulated samples. We show that two previously published defenses can be broken by better adapted attacks, highlighting the importance of robustness guarantees around OOD data. Since the existing method for this task is hard to train and significantly limits accuracy, we construct a classifier that can simultaneously achieve provability and high clean accuracy. Moreover, by architectural construction our method provably avoids the asymptotic overconfidence problem of standard neural networks.

## [Multi-Objective Deep Learning with Adaptive Reference Vectors](#)

- Weiyu Chen · James Kwok
- abstract@[open-review](#): Many deep learning models involve optimizing multiple objectives. Since objectives are often conflicting, we aim to get diverse and representative trade-off solutions among these objectives. Gradient-based multi-objective optimization (MOO) algorithms using reference vectors have shown promising performance. However, they may still produce undesirable solutions due to mismatch between the pre-specified reference vectors and the problem's underlying Pareto front. In this paper, we propose a novel gradient-based MOO algorithm with adaptive reference vectors. We formulate reference vector adaption as a bilevel optimization problem, and solve it with an efficient solver. Theoretical convergence analysis is also provided. Experiments on an extensive set of learning scenarios demonstrate the superiority of the proposed algorithm over the state-of-the-art.

## [Anonymous Bandits for Multi-User Systems](#)

- Hossein Esfandiari · Vahab Mirrokni · Jon Schneider
- abstract@[open-review](#): In this work, we present and study a new framework for online learning in systems with multiple users that provide user anonymity. Specifically, we extend the notion of bandits to obey the standard  $k$ -anonymity constraint by requiring each observation to be an aggregation of rewards for at least  $k$  users. This provides a simple yet effective framework where one can learn a clustering of users in an online fashion without observing any user's individual decision. We initiate the study of anonymous bandits and provide the first sublinear regret algorithms and lower bounds for this setting.

## [Benefits of Additive Noise in Composing Classes with Bounded Capacity](#)

- Alireza Fathollah Pour · Hassan Ashtiani
- abstract@[open-review](#): We observe that given two (compatible) classes of functions  $\mathcal{F}$  and  $\mathcal{H}$  with small capacity as measured by their uniform covering numbers, the capacity of the composition class  $\mathcal{H} \circ \mathcal{F}$  can become prohibitively large or even

unbounded. We then show that adding a small amount of Gaussian noise to the output of  $\mathcal{F}$  before composing it with  $\mathcal{H}$  can effectively control the capacity of  $\mathcal{H} \circ \mathcal{F}$ , offering a general recipe for modular design. To prove our results, we define new notions of uniform covering number of random functions with respect to the total variation and Wasserstein distances. We instantiate our results for the case of multi-layer neural networks. Preliminary empirical results indicate that the amount of noise required for our bound to improve over existing uniform bounds can be quite low.

## [Knowledge Distillation Improves Graph Structure Augmentation for Graph Neural Networks](#)

- Lirong Wu · Haitao Lin · Yufei Huang · Stan Z. Li
- abstract@[open-review](#): Graph (structure) augmentation aims to perturb the graph structure through heuristic or probabilistic rules, enabling the nodes to capture richer contextual information and thus improving generalization performance. While there have been a few graph structure augmentation methods proposed recently, none of them are aware of a potential negative augmentation problem, which may be caused by overly severe distribution shifts between the original and augmented graphs. In this paper, we take an important graph property, namely graph homophily, to analyze the distribution shifts between the two graphs and thus measure the severity of an augmentation algorithm suffering from negative augmentation. To tackle this problem, we propose a novel Knowledge Distillation for Graph Augmentation (KDGA) framework, which helps to reduce the potential negative effects of distribution shifts, i.e., negative augmentation problem. Specifically, KDGA extracts the knowledge of any GNN teacher model trained on the augmented graphs and injects it into a partially parameter-shared student model that is tested on the original graph. As a simple but efficient framework, KDGA is applicable to a variety of existing graph augmentation methods and can significantly improve the performance of various GNN architectures. For three popular graph augmentation methods, namely GAUG, MH-Aug, and GraphAug, the experimental results show that the learned student models outperform their vanilla implementations by an average accuracy of 4.6% (GAUG), 4.2% (MH-Aug), and 4.6% (GraphAug) on eight graph datasets.

## [Identifiability and generalizability from multiple experts in Inverse Reinforcement Learning](#)

- Paul Rolland · Luca Viano · Norman Schönhoff · Boris Nikolov · Volkan Cevher
- abstract@[open-review](#): While Reinforcement Learning (RL) aims to train an agent from a reward function in a given environment, Inverse Reinforcement Learning (IRL) seeks to recover the reward function from observing an expert's behavior. It is well known that, in general, various reward functions can lead to the same optimal policy, and hence, IRL is ill-defined. However, \cite{cao2021identifiability} showed that, if we observe two or more experts with different discount factors or acting in different environments, the reward function can under certain conditions be identified up to a constant. This work starts by showing an equivalent identifiability statement from multiple experts in tabular MDPs based on a rank condition, which is easily verifiable and is shown to be also necessary. We then extend our result to various different scenarios, i.e., we characterize reward identifiability in the case where the reward function can be represented as a linear combination of given features, making it more interpretable, or when we have access to approximate transition matrices. Even when the reward is not identifiable, we provide conditions characterizing when data on multiple experts in a given environment allows to generalize and train an optimal agent in a new environment. Our theoretical results on reward identifiability and generalizability are validated in various numerical experiments.

## [Learning Deep Input-Output Stable Dynamics](#)

- Ryosuke Kojima · Yuji Okamoto
- abstract@[open-review](#): Learning stable dynamics from observed time-series data is an essential problem in robotics, physical modeling, and systems biology. Many of these dynamics are represented as an inputs-output system to communicate with the external environment. In this study, we focus on input-output stable systems, exhibiting robustness against unexpected stimuli and noise. We propose a method to learn nonlinear systems guaranteeing the input-output stability. Our proposed method utilizes the differentiable projection onto the space satisfying the Hamilton-Jacobi inequality to realize the input-output stability. The problem of finding this projection can be formulated as a quadratic constraint quadratic programming problem, and we derive the particular solution analytically. Also, we apply our method to a toy bistable model and the task of training a benchmark generated from a glucose-insulin simulator. The results show that the nonlinear system with neural networks by our method achieves the input-output stability, unlike naive neural networks. Our code is available at <https://github.com/clinfo/DeepIOSStability>.

## [PyramidCLIP: Hierarchical Feature Alignment for Vision-language Model Pretraining](#)

- Yuting Gao · Jinfeng Liu · Zihan Xu · Jun Zhang · Ke Li · Rongrong Ji · Chunhua Shen
- abstract@[open-review](#): Large-scale vision-language pre-training has achieved promising results on downstream tasks. Existing methods highly rely on the assumption that the image-text pairs crawled from the Internet are in perfect one-to-one correspondence. However, in real scenarios, this assumption can be difficult to hold: the text description, obtained by crawling the affiliated metadata of the image, often suffers from the semantic mismatch and the mutual compatibility. To address these issues, we introduce PyramidCLIP, which constructs an input pyramid with different semantic levels for each modality, and aligns visual elements and linguistic elements in the form of hierarchy via peer-level semantics alignment and cross-level relation alignment. Furthermore, we soften the loss of negative samples (unpaired samples) so as to weaken the strict constraint during the pre-training stage, thus mitigating the risk of forcing the model to distinguish compatible negative pairs. Experiments on five downstream tasks demonstrate the effectiveness of the proposed PyramidCLIP. In particular, with the same amount of 15 million pre-training image-text pairs, PyramidCLIP exceeds CLIP on ImageNet zero-shot classification top-1 accuracy by 10.6%/13.2%/10.0% with ResNet50/ViT-B32/ViT-B16 based image encoder respectively. When scaling to larger datasets, PyramidCLIP achieves the state-of-the-art results on several downstream tasks. In particular, the results of PyramidCLIP-ResNet50 trained on 143M image-text pairs surpass that of CLIP using 400M data on ImageNet zero-shot classification task, significantly improving the data efficiency of CLIP.

## [Convexity Certificates from Hessians](#)

- Joachim Giesen · Julien Klaus · Sören Laue · Niklas Merk · Konstantin Wiedom
- abstract@[open-review](#): The Hessian of a differentiable convex function is positive semidefinite. Therefore, checking the Hessian of a given function is a natural approach to certify convexity. However, implementing this approach is not straightforward, since it requires a representation of the Hessian that allows its analysis. Here, we implement this approach for a class of functions that is rich enough to support classical machine learning. For this class of functions, it was recently shown how to compute computational graphs of their Hessians. We show how to check these graphs for positive-semidefiniteness. We compare our implementation of the Hessian approach with the well-established disciplined convex programming (DCP) approach and prove that the Hessian approach is at least as powerful as the DCP approach for differentiable functions. Furthermore, we show for a state-of-the-art implementation of the DCP approach that the Hessian approach is actually more powerful, that is, it can certify the convexity of a larger class of differentiable functions.

## [RISE: Robust Individualized Decision Learning with Sensitive Variables](#)

- Xiaoqing Tan · Zhengling Qi · Christopher Seymour · Lu Tang
- abstract@[open-review](#): This paper introduces RISE, a robust individualized decision learning framework with sensitive variables, where sensitive variables are collectible data and important to the intervention decision, but their inclusion in decision making is prohibited due to reasons such as delayed availability or fairness concerns. A naive baseline is to ignore these sensitive variables in learning decision rules, leading to significant uncertainty and bias. To address this, we propose a decision learning framework to incorporate sensitive variables during offline training but not include them in the input

of the learned decision rule during model deployment. Specifically, from a causal perspective, the proposed framework intends to improve the worst-case outcomes of individuals caused by sensitive variables that are unavailable at the time of decision. Unlike most existing literature that uses mean-optimal objectives, we propose a robust learning framework by finding a newly defined quantile- or infimum-optimal decision rule. The reliable performance of the proposed method is demonstrated through synthetic experiments and three real-data applications.

## [Neural Temporal Walks: Motif-Aware Representation Learning on Continuous-Time Dynamic Graphs](#)

- Ming Jin Â· Yuan-Fang Li Â· Shirui Pan
- abstract@[open-review](#): Continuous-time dynamic graphs naturally abstract many real-world systems, such as social and transactional networks. While the research on continuous-time dynamic graph representation learning has made significant advances recently, neither graph topological properties nor temporal dependencies have been well-considered and explicitly modeled in capturing dynamic patterns. In this paper, we introduce a novel method, Neural Temporal Walks (NeurTWs), for representation learning on continuous-time dynamic graphs. By considering not only time constraints but also structural and tree traversal properties, NeurTWs conducts spatiotemporal-biased random walks to retrieve a set of representative motifs, enabling temporal nodes to be characterized effectively. With a component based on neural ordinary differential equations, the extracted motifs allows for irregularly-sampled temporal nodes to be embedded explicitly over multiple interaction time intervals, enabling the capture of the underlying spatiotemporal dynamics. To enrich supervision signals, we further design a harder contrastive pretext task for model optimization. Our method demonstrates overwhelming superiority under both transductive and inductive settings on three real-world datasets.

## [Intrinsic dimensionality estimation using Normalizing Flows](#)

- Christian Horvat Â· Jean-Pascal Pfister
- abstract@[open-review](#): How many main invariances are there in a dataset consisting of  $M$  samples embedded in  $\mathbb{R}^D$ ? This number, formally known as  $\text{intrinsic dimensionality}$ , can be estimated using nearest neighbor statistics. However, nearest neighbor statistics do not scale to large datasets as their complexity scales quadratically in  $M$ ,  $\mathcal{O}(M^2)$ . Additionally, methods based on nearest neighbor statistics perform poorly on datasets embedded in high dimensions where  $D \gg 1$ . In this paper, we propose a novel method to estimate the intrinsic dimensionality using Normalizing Flows that scale to large datasets and high dimensions. Based on some back-of-the-envelope calculations, we predict how the eigenvalues of the flow's Jacobian evolve when inflating the dataset with different noise magnitudes. We test our method on various datasets, including 64x64 RGB images, where we achieve state-of-the-art results.

## [Learning Distributed and Fair Policies for Network Load Balancing as Markov Potential Game](#)

- Zhiyuan Yao Â· Zihan Ding
- abstract@[open-review](#): This paper investigates the network load balancing problem in data centers (DCs) where multiple load balancers (LBs) are deployed, using the multi-agent reinforcement learning (MARL) framework. The challenges of this problem consist of the heterogeneous processing architecture and dynamic environments, as well as limited and partial observability of each LB agent in distributed networking systems, which can largely degrade the performance of in-production load balancing algorithms in real-world setups. Centralised training and distributed execution (CTDE) RL scheme has been proposed to improve MARL performance, yet it incurs -- especially in distributed networking systems, which prefer distributed and plug-and-play design schemes -- additional communication and management overhead among agents. We formulate the multi-agent load balancing problem as a Markov potential game, with a carefully and properly designed workload distribution fairness as the potential function. A fully distributed MARL algorithm is proposed to approximate the Nash equilibrium of the game. Experimental evaluations involve both an event-driven simulator and a real-world system, where the proposed MARL load balancing algorithm shows close-to-optimal performance in simulations and superior results over in-production LBs in the real-world system.

## [Sleeper Agent: Scalable Hidden Trigger Backdoors for Neural Networks Trained from Scratch](#)

- Hossein Souri Â· Liam Fowl Â· Rama Chellappa Â· Micah Goldblum Â· Tom Goldstein
- abstract@[open-review](#): As the curation of data for machine learning becomes increasingly automated, dataset tampering is a mounting threat. Backdoor attackers tamper with training data to embed a vulnerability in models that are trained on that data. This vulnerability is then activated at inference time by placing a "trigger" into the model's input. Typical backdoor attacks insert the trigger directly into the training data, although the presence of such an attack may be visible upon inspection. In contrast, the Hidden Trigger Backdoor Attack achieves poisoning without placing a trigger into the training data at all. However, this hidden trigger attack is ineffective at poisoning neural networks trained from scratch. We develop a new hidden trigger attack, Sleeper Agent, which employs gradient matching, data selection, and target model re-training during the crafting process. Sleeper Agent is the first hidden trigger backdoor attack to be effective against neural networks trained from scratch. We demonstrate its effectiveness on ImageNet and in black-box settings.

## [Optimal Weak to Strong Learning](#)

- Kasper Green Larsen Â· Martin Ritzert
- abstract@[open-review](#): The classic algorithm AdaBoost allows to convert a weak learner, that is an algorithm that produces a hypothesis which is slightly better than chance, into a strong learner, achieving arbitrarily high accuracy when given enough training data. We present a new algorithm that constructs a strong learner from a weak learner, but uses less training data than AdaBoost and all other weak to strong learners to achieve the same generalization bounds. A sample complexity lower bound shows that our new algorithm uses the minimum possible amount of training data and is thus optimal. Hence, this work settles the sample complexity of the classic problem of constructing a strong learner from a weak learner.

## [Deformable Vision Transformer Based Single-Stage Pedestrian Detector](#)

- Jing Yuan Â· Panagiotis Barmpoutis Â· Tania Stathaki
- abstract@[open-review](#): Pedestrian detection is a challenging field in computer vision, which requires both fast inference and high accuracy. Single-stage detectors are faster than region of interest based two-stage detectors at the expense of accuracy mainly due to the lack of spatially adaptive features. To this end, we propose a single-stage anchor-free pedestrian detector with enhanced spatial and multi-scale features based on the deformable vision transformer aiming to achieve the balance between speed and accuracy. The design of the architecture is investigated in depth. Comprehensive comparisons with state-of-the-art single- and two- stage detectors on various pedestrian datasets are performed. The proposed detector achieves leading performance on both Caltech and Citypersons datasets among single- and two- stage methods using less parameters compared to the baseline. The log-average miss rates for Reasonable (3.8%) and Heavy (36.5%) are decreased to 2.6% and 28.0% on Caltech, and 10.6% and 36.7% on Citypersons validation datasets respectively. It even outperforms SOTA two-stage detectors in Heavy subset by 3% on Citypersons validation set.

## [On the Robustness of Graph Neural Diffusion](#)

- Yang Song Â· Qiyu Kang Â· Sijie Wang Â· Kai Zhao Â· Wee Peng Tay
- abstract@[open-review](#): Neural diffusion on graphs is a novel class of graph neural networks that has attracted increasing attention recently. The capability of graph neural partial differential equations (PDEs) in addressing common hurdles of graph neural networks (GNNs), such as the problems of over-smoothing and bottlenecks, has been investigated but not their robustness to adversarial attacks. In this work, we explore the robustness properties of

graph neural PDEs. We empirically demonstrate that graph neural PDEs are intrinsically more robust against topology perturbation as compared to other GNNs. We provide insights into this phenomenon by exploiting the stability of the heat semigroup under graph topology perturbations. We discuss various graph diffusion operators and relate them to existing graph neural PDEs. Furthermore, we propose a general graph neural PDE framework based on which a new class of robust GNNs can be defined. We verify that the new model achieves comparable state-of-the-art performance on several benchmark datasets.

## [Graph Neural Network Bandits](#)

- Parnian Kassraie · Andreas Krause · Ilija Bogunovic
- abstract@[open-review](#): We consider the bandit optimization problem with the reward function defined over graph-structured data. This problem has important applications in molecule design and drug discovery, where the reward is naturally invariant to graph permutations. The key challenges in this setting are scaling to large domains, and to graphs with many nodes. We resolve these challenges by embedding the permutation invariance into our model. In particular, we show that graph neural networks (GNNs) can be used to estimate the reward function, assuming it resides in the Reproducing Kernel Hilbert Space of a permutation-invariant additive kernel. By establishing a novel connection between such kernels and the graph neural tangent kernel (GNTK), we introduce the first GNN confidence bound and use it to design a phased-elimination algorithm with sublinear regret. Our regret bound depends on the GNTK's maximum information gain, which we also provide a bound for. Perhaps surprisingly, even though the reward function depends on all \$N\$ node features, our guarantees are independent of the number of graph nodes \$N\$. Empirically, our approach exhibits competitive performance and scales well on graph-structured domains.

## [What Makes Graph Neural Networks Miscalibrated?](#)

- Hans Hao-Hsun Hsu · Yuesong Shen · Christian Tomani · Daniel Cremers
- abstract@[open-review](#): Given the importance of getting calibrated predictions and reliable uncertainty estimations, various post-hoc calibration methods have been developed for neural networks on standard multi-class classification tasks. However, these methods are not well suited for calibrating graph neural networks (GNNs), which presents unique challenges such as the additional structural information and the graph-induced correlations between the nodes. In this work, we conduct a systematic study on the calibration qualities of GNN node predictions. And we identify five factors which influence the calibration of GNNs: general under-confident tendency, diversity of node distributions, distance to training nodes, relative confidence level, and neighborhood similarity. Furthermore, based on the insights from this study, we design a novel calibration method named Graph Attention Temperature Scaling (GATS), which is tailored for calibrating graph neural networks. GATS incorporates designs that address all the identified influential factors and produces nodewise temperature scaling using an attention-based architecture. GATS is accuracy-preserving, data-efficient, and expressive at the same time. Our experiments empirically verify the effectiveness of GATS, demonstrating that it can consistently achieve state-of-the-art calibration results on various graph datasets for different GNN backbones.

## [Neural network architecture beyond width and depth](#)

- Shijun Zhang · Zuowei Shen · Haizhao Yang
- abstract@[open-review](#): This paper proposes a new neural network architecture by introducing an additional dimension called height beyond width and depth. Neural network architectures with height, width, and depth as hyperparameters are called three-dimensional architectures. It is shown that neural networks with three-dimensional architectures are significantly more expressive than the ones with two-dimensional architectures (those with only width and depth as hyperparameters), e.g., standard fully connected networks. The new network architecture is constructed recursively via a nested structure, and hence we call a network with the new architecture nested network (NestNet). A NestNet of height \$s\$ is built with each hidden neuron activated by a NestNet of height \$\leq s-1\$. When \$s=1\$, a NestNet degenerates to a standard network with a two-dimensional architecture. It is proved by construction that height-\$s\$ ReLU NestNets with \$\mathcal{O}(n)\$ parameters can approximate Lipschitz continuous functions on \$[0,1]^d\$ with an error \$\mathcal{O}(n^{-(s+1)/d})\$, while the optimal approximation error of standard ReLU networks with \$\mathcal{O}(n)\$ parameters is \$\mathcal{O}(n^{-2/d})\$. Furthermore, such a result is extended to generic continuous functions on \$[0,1]^d\$ with the approximation error characterized by the modulus of continuity. Finally, a numerical example is provided to explore the advantages of the super approximation power of ReLU NestNets.

## [The Hessian Screening Rule](#)

- Johan Larsson · Jonas Wallin
- abstract@[open-review](#): Predictor screening rules, which discard predictors before fitting a model, have had considerable impact on the speed with which sparse regression problems, such as the lasso, can be solved. In this paper we present a new screening rule for solving the lasso path: the Hessian Screening Rule. The rule uses second-order information from the model to provide both effective screening, particularly in the case of high correlation, as well as accurate warm starts. The proposed rule outperforms all alternatives we study on simulated data sets with both low and high correlation for \$\ell\_1\$-regularized least-squares (the lasso) and logistic regression. It also performs best in general on the real data sets that we examine.

## [Graph Convolution Network based Recommender Systems: Learning Guarantee and Item Mixture Powered Strategy](#)

- Leyan Deng · Defu Lian · Chenwang Wu · Enhong Chen
- abstract@[open-review](#): Inspired by their powerful representation ability on graph-structured data, Graph Convolution Networks (GCNs) have been widely applied to recommender systems, and have shown superior performance. Despite their empirical success, there is a lack of theoretical explorations such as generalization properties. In this paper, we take a first step towards establishing a generalization guarantee for GCN-based recommendation models under inductive and transductive learning. We mainly investigate the roles of graph normalization and non-linear activation, providing some theoretical understanding, and construct extensive experiments to further verify these findings empirically. Furthermore, based on the proven generalization bound and the challenge of existing models in discrete data learning, we propose Item Mixture (IMix) to enhance recommendation. It models discrete spaces in a continuous manner by mixing the embeddings of positive-negative item pairs, and its effectiveness can be strictly guaranteed from empirical and theoretical aspects.

## [Revisiting Neural Scaling Laws in Language and Vision](#)

- Ibrahim Alabdulmohsin · Behnam Neyshabur · Xiaohua Zhai
- abstract@[open-review](#): The remarkable progress in deep learning in recent years is largely driven by improvements in scale, where bigger models are trained on larger datasets for longer schedules. To predict the benefit of scale empirically, we argue for a more rigorous methodology based on the extrapolation loss, instead of reporting the best-fitting (interpolating) parameters. We then present a recipe for estimating scaling law parameters reliably from learning curves. We demonstrate that it extrapolates more accurately than previous methods in a wide range of architecture families across several domains, including image classification, neural machine translation (NMT) and language modeling, in addition to tasks from the BIG-Bench evaluation benchmark. Finally, we release a benchmark dataset comprising of 90 evaluation tasks to facilitate research in this domain.

## [Average Sensitivity of Euclidean k-Clustering](#)

- Yuichi Yoshida · Shinji Ito

- abstract@[open-review](#): Given a set of  $n$  points in  $\mathbb{R}^d$ , the goal of Euclidean  $(k, \ell)$ -clustering is to find  $k$  centers that minimize the sum of the  $\ell$ -th powers of the Euclidean distance of each point to the closest center. In practical situations, the clustering result must be stable against points missing in the input data so that we can make trustworthy and consistent decisions. To address this issue, we consider the average sensitivity of Euclidean  $(k, \ell)$ -clustering, which measures the stability of the output in total variation distance against deleting a random point from the input data. We first show that a popular algorithm `k-means++` and its variant called `Dell-sampling` have low average sensitivity. Next, we show that any approximation algorithm for Euclidean  $(k, \ell)$ -clustering can be transformed to an algorithm with low average sensitivity while almost preserving the approximation guarantee. As byproducts of our results, we provide several algorithms for consistent  $(k, \ell)$ -clustering and dynamic  $(k, \ell)$ -clustering in the random-order model, where the input points are randomly permuted and given in an online manner. The goal of the consistent setting is to maintain a good solution while minimizing the number of changes to the solution during the process, and that of the dynamic setting is to maintain a good solution while minimizing the (amortized) update time.

## [A Reduction to Binary Approach for Debiasing Multiclass Datasets](#)

- Ibrahim Alabdulmohsin · Jessica Schrouff · Sanmi Koyejo
- abstract@[open-review](#): We propose a novel reduction-to-binary (R2B) approach that enforces demographic parity for multiclass classification with non-binary sensitive attributes via a reduction to a sequence of binary debiasing tasks. We prove that R2B satisfies optimality and bias guarantees and demonstrate empirically that it can lead to an improvement over two baselines: (1) treating multiclass problems as multi-label by debiasing labels independently and (2) transforming the features instead of the labels. Surprisingly, we also demonstrate that independent label debiasing yields competitive results in most (but not all) settings. We validate these conclusions on synthetic and real-world datasets from social science, computer vision, and healthcare.

## [A Solver-free Framework for Scalable Learning in Neural ILP Architectures](#)

- Yatin Nandwani · Rishabh Ranjan · Mausam · Parag Singla
- abstract@[open-review](#): There is a recent focus on designing architectures that have an Integer Linear Programming (ILP) layer within a neural model (referred to as `Neural ILP` in this paper). Neural ILP architectures are suitable for pure reasoning tasks that require data-driven constraint learning or for tasks requiring both perception (neural) and reasoning (ILP). A recent SOTA approach for end-to-end training of Neural ILP explicitly defines gradients through the ILP black box [Paulus et al. [2021]] — this trains extremely slowly, owing to a call to the underlying ILP solver for every training data point in a minibatch. In response, we present an alternative training strategy that is `solver-free`, i.e., does not call the ILP solver at all at training time. Neural ILP has a set of trainable hyperplanes (for cost and constraints in ILP), together representing a polyhedron. Our key idea is that the training loss should impose that the final polyhedron separates the positives (all constraints satisfied) from the negatives (at least one violated constraint or a suboptimal cost value), via a soft-margin formulation. While positive example(s) are provided as part of the training data, we devise novel techniques for generating negative samples. Our solution is flexible enough to handle equality as well as inequality constraints. Experiments on several problems, both perceptual as well as symbolic, which require learning the constraints of an ILP, show that our approach has superior performance and scales much better compared to purely neural baselines and other state-of-the-art models that require solver-based training. In particular, we are able to obtain excellent performance in 9 x 9 Visual Sudoku, to which the other Neural ILP solver is not able to scale.

## [Combinatorial Bandits with Linear Constraints: Beyond Knapsacks and Fairness](#)

- Qingsong Liu · Weihang Xu · Siwei Wang · Zhixuan Fang
- abstract@[open-review](#): This paper proposes and studies for the first time the problem of combinatorial multi-armed bandits with linear long-term constraints. Our model generalizes and unifies several prominent lines of work, including bandits with fairness constraints, bandits with knapsacks (BwK), etc. We propose an upper-confidence bound LP-style algorithm for this problem, called UCB-LP, and prove that it achieves a logarithmic problem-dependent regret bound and zero constraint violations in expectation. In the special case of fairness constraints, we further provide a sharper constant regret bound for UCB-LP. Our regret bounds outperform the existing literature on BwK and bandits with fairness constraints simultaneously. We also develop another low-complexity version of UCB-LP and show that it yields  $O(\sqrt{T})$  problem-independent regret and zero constraint violations with high-probability. Finally, we conduct numerical experiments to validate our theoretical results.

## [Unifying and Boosting Gradient-Based Training-Free Neural Architecture Search](#)

- YAO SHU · Zhongxiang Dai · Zhaoxuan Wu · Bryan Kian Hsiang Low
- abstract@[open-review](#): Neural architecture search (NAS) has gained immense popularity owing to its ability to automate neural architecture design. A number of training-free metrics are recently proposed to realize NAS without training, hence making NAS more scalable. Despite their competitive empirical performances, a unified theoretical understanding of these training-free metrics is lacking. As a consequence, (a) the relationships among these metrics are unclear, (b) there is no theoretical interpretation for their empirical performances, and (c) there may exist untapped potential in existing training-free NAS, which probably can be unveiled through a unified theoretical understanding. To this end, this paper presents a unified theoretical analysis of gradient-based training-free NAS, which allows us to (a) theoretically study their relationships, (b) theoretically guarantee their generalization performances, and (c) exploit our unified theoretical understanding to develop a novel framework named hybrid NAS (HNAS) which consistently boosts training-free NAS in a principled way. Remarkably, HNAS can enjoy the advantages of both training-free (i.e., superior search efficiency) and training-based (i.e., remarkable search effectiveness) NAS, which we have demonstrated through extensive experiments.

## [Robust Feature-Level Adversaries are Interpretability Tools](#)

- Stephen Casper · Max Nadeau · Dylan Hadfield-Menell · Gabriel Kreiman
- abstract@[open-review](#): The literature on adversarial attacks in computer vision typically focuses on pixel-level perturbations which tend to be very difficult to interpret. Recent work that manipulates the latent representations of image generators to create "feature-level" adversarial perturbations gives us an opportunity to explore perceptible, interpretable adversarial attacks. We make three contributions. First, we observe that feature-level attacks provide useful classes of inputs for studying the representations in models. Second, we show that these adversaries are versatile and highly robust. We demonstrate that they can be used to produce targeted, universal, disguised, physically-realizable, and black-box attacks at the ImageNet scale. Third, we show how these adversarial images can be used as a practical interpretability tool for identifying bugs in networks. We use these adversaries to make predictions about spurious associations between features and classes which we then test by designing "copy/paste" attacks in which one natural image is pasted into another to cause a targeted misclassification. Our results indicate that feature-level attacks are a promising approach for rigorous interpretability research. They support the design of tools to better understand what a model has learned and diagnose brittle feature associations.

## [Concrete Score Matching: Generalized Score Matching for Discrete Data](#)

- Chenlin Meng · Kristy Choi · Jiaming Song · Stefano Ermon
- abstract@[open-review](#): Representing probability distributions by the gradient of their density functions has proven effective in modeling a wide range of continuous data modalities. However, this representation is not applicable in discrete domains where the gradient is undefined. To this end, we propose an analogous score function called the "Concrete score", a generalization of the (Stein) score for discrete settings. Given a predefined neighborhood structure, the Concrete score of any input is defined by the rate of change of the probabilities with respect to local directional changes of the input. This formulation allows us to recover the (Stein) score in continuous domains when measuring such changes by the Euclidean distance, while using the

Manhattan distance leads to our novel score function in discrete domains. Finally, we introduce a new framework to learn such scores from samples called Concrete Score Matching (CSM), and propose an efficient training objective to scale our approach to high dimensions. Empirically, we demonstrate the efficacy of CSM on density estimation tasks on a mixture of synthetic, tabular, and high-dimensional image datasets, and demonstrate that it performs favorably relative to existing baselines for modeling discrete data.

## [On Elimination Strategies for Bandit Fixed-Confidence Identification](#)

- Andrea Tirinzoni · Romain Degenne
- abstract@[open-review](#): Elimination algorithms for bandit identification, which prune the plausible correct answers sequentially until only one remains, are computationally convenient since they reduce the problem size over time. However, existing elimination strategies are often not fully adaptive (they update their sampling rule infrequently) and are not easy to extend to combinatorial settings, where the set of answers is exponentially large in the problem dimension. On the other hand, most existing fully-adaptive strategies to tackle general identification problems are computationally demanding since they repeatedly test the correctness of every answer, without ever reducing the problem size. We show that adaptive methods can be modified to use elimination in both their stopping and sampling rules, hence obtaining the best of these two worlds: the algorithms (1) remain fully adaptive, (2) suffer a sample complexity that is never worse of their non-elimination counterpart, and (3) provably eliminate certain wrong answers early. We confirm these benefits experimentally, where elimination improves significantly the computational complexity of adaptive methods on common tasks like best-arm identification in linear bandits.

## [The Mechanism of Prediction Head in Non-contrastive Self-supervised Learning](#)

- Zixin Wen · Yuanzhi Li
- abstract@[open-review](#): The surprising discovery of the BYOL method shows the negative samples can be replaced by adding the prediction head to the network. It is mysterious why even when there exist trivial collapsed global optimal solutions, neural networks trained by (stochastic) gradient descent can still learn competitive representations. In this work, we present our empirical and theoretical discoveries on non-contrastive self-supervised learning. Empirically, we find that when the prediction head is initialized as an identity matrix with only its off-diagonal entries being trainable, the network can learn competitive representations even though the trivial optima still exist in the training objective. Theoretically, we characterized the substitution effect and acceleration effect of the trainable, but identity-initialized prediction head. The substitution effect happens when learning the stronger features in some neurons can substitute for learning these features in other neurons through updating the prediction head. And the acceleration effect happens when the substituted features can accelerate the learning of other weaker features to prevent them from being ignored. These two effects enable the neural networks to learn diversified features rather than focus only on learning the strongest features, which is likely the cause of the dimensional collapse phenomenon. To the best of our knowledge, this is also the first end-to-end optimization guarantee for non-contrastive methods using nonlinear neural networks with a trainable prediction head and normalization.

## [Quantum Speedups of Optimizing Approximately Convex Functions with Applications to Logarithmic Regret Stochastic Convex Bandits](#)

- Tongyang Li · Ruizhe Zhang
- abstract@[open-review](#): We initiate the study of quantum algorithms for optimizing approximately convex functions. Given a convex set  $\mathcal{K} \subseteq \mathbb{R}^n$  and a function  $F: \mathbb{R}^n \rightarrow \mathbb{R}$  satisfying  $\sup_{x \in \mathcal{K}} |F(x) - f(x)| \leq \epsilon/n$ , our quantum algorithm finds an  $x \in \mathcal{K}$  such that  $|F(x) - \min_{x \in \mathcal{K}} F(x)| \leq \epsilon$  using  $\tilde{O}(n^3)$  quantum evaluation queries to  $F$ . This achieves a polynomial quantum speedup compared to the best-known classical algorithms. As an application, we give a quantum algorithm for zeroth-order stochastic convex bandits with  $\tilde{O}(n^5 \log^2 T)$  regret, an exponential speedup in  $T$  compared to the classical  $\Omega(\sqrt{T})$  lower bound. Technically, we achieve quantum speedup in  $n$  by exploiting a quantum framework of simulated annealing and adopting a quantum version of the hit-and-run walk. Our speedup in  $T$  for zeroth-order stochastic convex bandits is due to a quadratic quantum speedup in multiplicative error of mean estimation.

## [Exploiting the Relationship Between Kendall's Rank Correlation and Cosine Similarity for Attribution Protection](#)

- Fan Wang · Adams Wai Kin Kong
- abstract@[open-review](#): Model attributions are important in deep neural networks as they aid practitioners in understanding the models, but recent studies reveal that attributions can be easily perturbed by adding imperceptible noise to the input. The non-differentiable Kendall's rank correlation is a key performance index for attribution protection. In this paper, we first show that the expected Kendall's rank correlation is positively correlated to cosine similarity and then indicate that the direction of attribution is the key to attribution robustness. Based on these findings, we explore the vector space of attribution to explain the shortcomings of attribution defense methods using  $\ell_p$  norm and propose integrated gradient regularizer (IGR), which maximizes the cosine similarity between natural and perturbed attributions. Our analysis further exposes that IGR encourages neurons with the same activation states for natural samples and the corresponding perturbed samples. Our experiments on different models and datasets confirm our analysis on attribution protection and demonstrate a decent improvement in adversarial robustness.

## [Learning to Mitigate AI Collusion on Economic Platforms](#)

- Eric Mibuary · Gianluca Brero · David Parkes · Nicolas Lepore
- abstract@[open-review](#): Algorithmic pricing on online e-commerce platforms raises the concern of tacit collusion, where reinforcement learning algorithms learn to set collusive prices in a decentralized manner and through nothing more than profit feedback. This raises the question as to whether collusive pricing can be prevented through the design of suitable "buy boxes," i.e., through the design of the rules that govern the elements of e-commerce sites that promote particular products and prices to consumers. In this paper, we demonstrate that reinforcement learning (RL) can also be used by platforms to learn buy box rules that are effective in preventing collusion by RL sellers. For this, we adopt the methodology of Stackelberg POMDPs, and demonstrate success in learning robust rules that continue to provide high consumer welfare together with sellers employing different behavior models or having out-of-distribution costs for goods.

## [â€œWhy Not Other Classes?â€ : Towards Class-Contrastive Back-Propagation Explanations](#)

- Yipei Wang · Xiaoqian Wang
- abstract@[open-review](#): Numerous explanation methods have been developed for deep neural networks (DNNs) based classifiers. The goal of these methods is to explore the inner mechanism of DNNs. Existing explanation methods are often limited to explaining predictions of a pre-specified class, which answers the question "why is the input classified into this class?" However, such explanations with respect to a single class are inherently insufficient because they do not capture features with class-discriminative power. That is, features that are important for predicting one class may also be important for other classes. To capture features with true class-discriminative power, we should instead ask "why is the input classified into this class, but not others?" To answer this question, we propose a weighted contrastive framework for explaining DNNs. Our framework can easily convert any existing back-propagation explanation methods to build class-contrastive explanations. We theoretically validate our weighted contrast explanation in general back-propagation explanations for DNNs. And we validate that our method effectively enables class-contrastive explanations with significant improvements in

both qualitative and quantitative experiments. Based on the theoretical and experimental results, we suggest that contrastive explanations are necessary and should always be considered in explaining classification models.

## [Self-Supervised Pretraining for Large-Scale Point Clouds](#)

- Zaiwei Zhang · Min Bai · Li Erran Li
- abstract@[open-review](#): Pretraining on large unlabeled datasets has been proven to improve the down-stream task performance on many computer vision tasks, such as 2D object detection and video classification. However, for large-scale 3D scenes, such as outdoor LiDAR point clouds, pretraining is not widely used. Due to the special data characteristics of large 3D point clouds, 2D pretraining frameworks tend to not generalize well. In this paper, we propose a new self-supervised pretraining method that targets large-scale 3D scenes. We pretrain commonly used point-based and voxel-based model architectures and show the transfer learning performance on 3D object detection and also semantic segmentation. We demonstrate the effectiveness of our approach on both dense 3D indoor point clouds and also sparse outdoor lidar point clouds.

## [Lower Bounds and Nearly Optimal Algorithms in Distributed Learning with Communication Compression](#)

- Xinmeng Huang · Yiming Chen · Wotao Yin · Kun Yuan
- abstract@[open-review](#): Recent advances in distributed optimization and learning have shown that communication compression is one of the most effective means of reducing communication. While there have been many results for convergence rates with compressed communication, a lower bound is still missing. Analyses of algorithms with communication compression have identified two abstract properties that guarantee convergence: the unbiased property or the contractive property. They can be applied either unidirectionally (compressing messages from worker to server) or bidirectionally. In the smooth and non-convex stochastic regime, this paper establishes a lower bound for distributed algorithms whether using unbiased or contractive compressors in unidirection or bidirection. To close the gap between this lower bound and the best existing upper bound, we further propose an algorithm, NEOLITHIC, that almost reaches our lower bound (except for a logarithm factor) under mild conditions. Our results also show that using contractive compressors in bidirection can yield iterative methods that converge as fast as those using unbiased compressors unidirectionally. We report experimental results that validate our findings.

## [Continuous Deep Q-Learning in Optimal Control Problems: Normalized Advantage Functions Analysis](#)

- Anton Plaksin · Stepan Martyanov
- abstract@[open-review](#): One of the most effective continuous deep reinforcement learning algorithms is normalized advantage functions (NAF). The main idea of NAF consists in the approximation of the Q-function by functions quadratic with respect to the action variable. This idea allows to apply the algorithm to continuous reinforcement learning problems, but on the other hand, it brings up the question of classes of problems in which this approximation is acceptable. The presented paper describes one such class. We consider reinforcement learning problems obtained by the discretization of certain optimal control problems. Based on the idea of NAF, we present a new family of quadratic functions and prove its suitable approximation properties. Taking these properties into account, we provide several ways to improve NAF. The experimental results confirm the efficiency of our improvements.

## [Distributed Distributionally Robust Optimization with Non-Convex Objectives](#)

- Yang Jiao · Kai Yang · Dongjin Song
- abstract@[open-review](#): Distributionally Robust Optimization (DRO), which aims to find an optimal decision that minimizes the worst case cost over the ambiguity set of probability distribution, has been applied in diverse applications, e.g., network behavior analysis, risk management, etc. However, existing DRO techniques face three key challenges: 1) how to deal with the asynchronous updating in a distributed environment; 2) how to leverage the prior distribution effectively; 3) how to properly adjust the degree of robustness according to difference scenarios. To this end, we propose an asynchronous distributed algorithm, named Asynchronous Single-loop alternatIve gRadient projEction (ASPIRE) algorithm with the itErative Active SEt method (EASE) to tackle the distributed distributionally robust optimization (DDRO) problem. Furthermore, a new uncertainty set, i.e., constrained  $\$D\$$ -norm uncertainty set, is developed to effectively leverage the prior distribution and flexibly control the degree of robustness. Finally, our theoretical analysis elucidates that the proposed algorithm is guaranteed to converge and the iteration complexity is also analyzed. Extensive empirical studies on real-world datasets demonstrate that the proposed method can not only achieve fast convergence, remain robust against data heterogeneity and malicious attacks, but also tradeoff robustness with performance.

## [VF-PS: How to Select Important Participants in Vertical Federated Learning, Efficiently and Securely?](#)

- Jiawei Jiang · Lukas Burkhalter · Fangcheng Fu · Bolin Ding · Bo Du · Anwar Hithnawi · Bo Li · Ce Zhang
- abstract@[open-review](#): Vertical Federated Learning (VFL), that trains federated models over vertically partitioned data, has emerged as an important learning paradigm. However, existing VFL methods are facing two challenges: (1) scalability when # participants grows to even modest scale and (2) diminishing return w.r.t. # participants: not all participants are equally important and many will not introduce quality improvement in a large consortium. Inspired by these two challenges, in this paper, we ask: How can we select  $\$l\$$  out of  $\$m\$$  participants, where  $\$l \ll m\$$ , that are most important? We call this problem Vertically Federated Participant Selection, and model it with a principled mutual information-based view. Our first technical contribution is VF-MINE---a Vertically Federated Mutual INformation ESTimator---that uses one of the most celebrated algorithms in database theory---Fagin's algorithm as a building block. Our second contribution is to further optimize VF-MINE to enable VF-PS, a group testing-based participant selection framework. We empirically show that vertically federated participation selection can be orders of magnitude faster than training a full-fledged VFL model, while being able to identify the most important subset of participants that often lead to a VFL model of similar quality.

## [Unsupervised Image-to-Image Translation with Density Changing Regularization](#)

- Shaoan Xie · Qirong Ho · Kun Zhang
- abstract@[open-review](#): Unpaired image-to-image translation aims to translate an input image to another domain such that the output image looks like an image from another domain while important semantic information are preserved. Inferring the optimal mapping with unpaired data is impossible without making any assumptions. In this paper, we make a density changing assumption where image patches of high probability density should be mapped to patches of high probability density in another domain. Then we propose an efficient way to enforce this assumption: we train the flows as density estimators and penalize the variance of density changes. Despite its simplicity, our method achieves the best performance on benchmark datasets and needs only  $\$56\%-86\%\$$  of training time of the existing state-of-the-art method.

## [Quantile Constrained Reinforcement Learning: A Reinforcement Learning Framework Constraining Outage Probability](#)

- Whiyoung Jung · Myungsik Cho · Jongeui Park · Youngchul Sung
- abstract@[open-review](#): Constrained reinforcement learning (RL) is an area of RL whose objective is to find an optimal policy that maximizes expected cumulative return while satisfying a given constraint. Most of the previous constrained RL works consider expected cumulative sum cost as the constraint. However, optimization with this constraint cannot guarantee a target probability of outage event that the cumulative sum cost exceeds a given threshold. This paper proposes a framework, named Quantile Constrained RL (QCRL), to constrain the quantile of the distribution of cumulative sum cost that is a

necessary and sufficient condition to satisfy the outage constraint. This is the first work that tackles the issue of applying the policy gradient theorem to the quantile and provides theoretical results for approximating the gradient of the quantile. Based on the derived theoretical results and the technique of the Lagrange multiplier, we construct a constrained RL algorithm named Quantile Constrained Policy Optimization (QCPO). We use distributional RL with the Large Deviation Principle (LDP) to estimate quantiles and tail probability of cumulative sum cost for the implementation of QCPO. The implemented algorithm satisfies the outage probability constraint after the training period.

## [Does Self-supervised Learning Really Improve Reinforcement Learning from Pixels?](#)

- Xiang Li · Jinghuan Shang · Srijan Das · Michael Ryoo
- abstract@[open-review](#): We investigate whether self-supervised learning (SSL) can improve online reinforcement learning (RL) from pixels. We extend the contrastive reinforcement learning framework (e.g., CURL) that jointly optimizes SSL and RL losses and conduct an extensive amount of experiments with various self-supervised losses. Our observations suggest that the existing SSL framework for RL fails to bring meaningful improvement over the baselines only taking advantage of image augmentation when the same amount of data and augmentation is used. We further perform an evolutionary search to find the optimal combination of multiple self-supervised losses for RL, but find that even such a loss combination fails to meaningfully outperform the methods that only utilize carefully designed image augmentations. Often, the use of self-supervised losses under the existing framework lowered RL performances. We evaluate the approach in multiple different environments including a real-world robot environment and confirm that no single self-supervised loss or image augmentation method can dominate all environments and that the current framework for joint optimization of SSL and RL is limited. Finally, we conduct the ablation study on multiple factors and demonstrate the properties of representations learned with different approaches.

## [An Online Algorithm for Data Deletion](#)

- Vinith Suriyakumar · Ashia Wilson
- abstract@[open-review](#): We study the problem of deleting user data from machine learning models trained using empirical risk minimization (ERM). Our focus is on learning algorithms which return the empirical risk minimizer and approximate unlearning algorithms that comply with deletion requests that come in an online manner. Leveraging the infinitesimal jackknife, we develop an online unlearning algorithm that is both computationally and memory efficient. Unlike prior memory efficient unlearning algorithms, we target ERM trained models that minimize objectives with non-smooth regularizers, such as the commonly used  $\ell_1$ , elastic net, or nuclear norm penalties. We also provide generalization, deletion capacity, and unlearning guarantees that are consistent with state of the art methods. Across a variety of benchmark datasets, our algorithm empirically improves upon the runtime of prior methods while maintaining the same memory requirements and test accuracy. Finally, we open a new direction of inquiry by proving that all approximate unlearning algorithms introduced so far fail to unlearn in problem settings where common hyperparameter tuning methods, such as cross-validation, have been used to select models.

## [Generative Neural Articulated Radiance Fields](#)

- Alexander Bergman · Petr Kellnhofer · Wang Yifan · Eric Chan · David Lindell · Gordon Wetzstein
- abstract@[open-review](#): Unsupervised learning of 3D-aware generative adversarial networks (GANs) using only collections of single-view 2D photographs has very recently made much progress. These 3D GANs, however, have not been demonstrated for human bodies and the generated radiance fields of existing frameworks are not directly editable, limiting their applicability in downstream tasks. We propose a solution to these challenges by developing a 3D GAN framework that learns to generate radiance fields of human bodies or faces in a canonical pose and warp them using an explicit deformation field into a desired body pose or facial expression. Using our framework, we demonstrate the first high-quality radiance field generation results for human bodies. Moreover, we show that our deformation-aware training procedure significantly improves the quality of generated bodies or faces when editing their poses or facial expressions compared to a 3D GAN that is not trained with explicit deformations.

## [Exposing and Exploiting Fine-Grained Block Structures for Fast and Accurate Sparse Training](#)

- Peng Jiang · Lihuan Hu · Shihui Song
- abstract@[open-review](#): Sparse training is a popular technique to reduce the overhead of training large models. Although previous work has shown promising results for nonstructured sparse models, it is still unclear whether a sparse model with structural constraints can be trained from scratch to high accuracy. In this work, we study the dynamic sparse training for a class of sparse models with shuffled block structures. Compared to nonstructured models, such fine-grained structured models are more hardware-friendly and can effectively accelerate the training process. We propose an algorithm that keeps adapting the sparse model while maintaining the active parameters in shuffled blocks. We conduct experiments on a variety of networks and datasets and obtain positive results. In particular, on ImageNet, we achieve dense accuracy for ResNet50 and ResNet18 at 0.5 sparsity. On CIFAR10/100, we show that dense accuracy can be recovered at 0.6 sparsity for various models. At higher sparsity, our algorithm can still match the accuracy of nonstructured sparse training in most cases, while reducing the training time by up to 5x due to the fine-grained block structures in the models.

## [Non-Linguistic Supervision for Contrastive Learning of Sentence Embeddings](#)

- Yiren Jian · Chongyang Gao · Soroush Vosoughi
- abstract@[open-review](#): Semantic representation learning for sentences is an important and well-studied problem in NLP. The current trend for this task involves training a Transformer-based sentence encoder through a contrastive objective with text, i.e., clustering sentences with semantically similar meanings and scattering others. In this work, we find the performance of Transformer models as sentence encoders can be improved by training with multi-modal multi-task losses, using unpaired examples from another modality (e.g., sentences and unrelated image/audio data). In particular, besides learning by the contrastive loss on text, our model clusters examples from a non-linguistic domain (e.g., visual/audio) with a similar contrastive loss at the same time. The reliance of our framework on unpaired non-linguistic data makes it language-agnostic, enabling it to be widely applicable beyond English NLP. Experiments on 7 semantic textual similarity benchmarks reveal that models trained with the additional non-linguistic (images/audio) contrastive objective lead to higher quality sentence embeddings. This indicates that Transformer models are able to generalize better by doing a similar task (i.e., clustering) with \textit{unpaired} examples from different modalities in a multi-task fashion. The code is available at <https://github.com/yiren-jian/NonLing-CSE>.

## [A Combinatorial Perspective on the Optimization of Shallow ReLU Networks](#)

- Michael S Matena · Colin Raffel
- abstract@[open-review](#): The NP-hard problem of optimizing a shallow ReLU network can be characterized as a combinatorial search over the activation pattern for each training example followed by a constrained convex problem given a fixed set of activation patterns. We explore the implications of this combinatorial aspect of ReLU optimization in this work. We show that it can be naturally modeled via a geometric and combinatoric object known as a zonotope with its vertex set isomorphic to the set of feasible activation patterns. This assists in analysis and provides a foundation for further research. We provide an example of its usefulness when we explore the sensitivity of the optimal loss to perturbations of the training data. Later we discuss methods of zonotope vertex selection and its relevance to optimization. Overparameterization assists in training by making a randomly chosen vertex more likely to contain a good solution. We then introduce a novel polynomial-time vertex selection procedure that provably picks a vertex containing the global optimum using only double the minimum number of parameters required to fit the data. We further introduce a local greedy search heuristic over zonotope vertices and demonstrate that it outperforms gradient descent on underparameterized problems.

## [Physics-Informed Implicit Representations of Network Flows](#)

- Kevin D. Smith · Francesco Seccamonte · Ananthram Swami · Francesco Bullo
- abstract@[open-review](#): Flow networks are ubiquitous in natural and engineered systems, and in order to understand and manage these networks, one must quantify the flow of commodities across their edges. This paper considers the estimation problem of predicting unlabeled edge flows from nodal supply and demand. We propose an implicit neural network layer that incorporates two fundamental physical laws: conservation of mass, and the existence of a constitutive relationship between edge flows and nodal states (e.g., Ohm's law). Computing the edge flows from these two laws is a nonlinear inverse problem, which our layer solves efficiently with a specialized contraction mapping. Using implicit differentiation to compute the solution's gradients, our model is able to learn the constitutive relationship within a semi-supervised framework. We demonstrate that our approach can accurately predict edge flows in several experiments on AC power networks and water distribution systems.

## [On the Discrimination Risk of Mean Aggregation Feature Imputation in Graphs](#)

- Arjun Subramonian · Kai-Wei Chang · Yizhou Sun
- abstract@[open-review](#): While machine learning methods have demonstrated success on graph-structured data, many methods rely on fully-observed node features. This has led to an increase in research on imputing unknown or missing node features. However, in human networks, nodes belonging to a marginalized group have a disproportionate rate of unknown features. Human networks also contain graph structure and known feature biases. All these factors can cause graph feature imputation algorithms to predict values for unknown features that cause the feature values of the marginalized group to be more distinct on average from the feature values of the dominant group than they are in reality. We call this distinction the discrimination risk. We prove that a higher discrimination risk can amplify the unfairness of a machine learning model trained on the imputed data. We then formalize a general graph feature imputation framework called mean aggregation imputation and theoretically and empirically characterize graphs in which applying this framework can yield imputed features with a high discrimination risk. We propose a simple and effective solution to ensure mean aggregation-imputed features provably have a low discrimination risk (while minimally sacrificing utility) and improve the fairness of models. We evaluate the fairness and utility of our solution on synthetic and real-world credit network datasets.

## [Towards Understanding the Condensation of Neural Networks at Initial Training](#)

- Hanxu Zhou · Zhou Qixuan · Tao Luo · Yaoyu Zhang · Zhi-Qin Xu
- abstract@[open-review](#): Empirical works show that for ReLU neural networks (NNs) with small initialization, input weights of hidden neurons (the input weight of a hidden neuron consists of the weight from its input layer to the hidden neuron and its bias term) condense onto isolated orientations. The condensation dynamics implies that the training implicitly regularizes a NN towards one with much smaller effective size. In this work, we illustrate the formation of the condensation in multi-layer fully connected NNs and show that the maximal number of condensed orientations in the initial training stage is twice the multiplicity of the activation function, where ``multiplicity'' indicates the multiple roots of activation function at origin. Our theoretical analysis confirms experiments for two cases, one is for the activation function of multiplicity one with arbitrary dimension input, which contains many common activation functions, and the other is for the layer with one-dimensional input and arbitrary multiplicity. This work makes a step towards understanding how small initialization leads NNs to condensation at the initial training stage.

## [Rapidly Mixing Multiple-try Metropolis Algorithms for Model Selection Problems](#)

- Hyunwoong Chang · Changwoo Lee · Zhao Tang Luo · Huiyan Sang · Quan Zhou
- abstract@[open-review](#): The Multiple-try Metropolis (MTM) algorithm is an extension of the Metropolis-Hastings (MH) algorithm by selecting the proposed state among multiple trials according to some weight function. Although MTM has gained great popularity owing to its faster empirical convergence and mixing than the standard MH algorithm, its theoretical mixing property is rarely studied in the literature due to its complex proposal scheme. We prove that MTM can achieve a mixing time bound smaller than that of MH by a factor of the number of trials under a general setting applicable to high-dimensional model selection problems. Our theoretical results motivate a new class of weight functions and guide the choice of the number of trials, which leads to improved performance than standard MTM algorithms. We support our theoretical results by extensive simulation studies with several Bayesian model selection problems: variable selection, stochastic block models, and spatial clustering models.

## [Emergent Graphical Conventions in a Visual Communication Game](#)

- Shuwen Qiu · Sirui Xie · Lifeng Fan · Tao Gao · Jungseock Joo · Song-Chun Zhu · Yixin Zhu
- abstract@[open-review](#): Humans communicate with graphical sketches apart from symbolic languages. Primarily focusing on the latter, recent studies of emergent communication overlook the sketches; they do not account for the evolution process through which symbolic sign systems emerge in the trade-off between iconicity and symbolicity. In this work, we take the very first step to model and simulate this process via two neural agents playing a visual communication game; the sender communicates with the receiver by sketching on a canvas. We devise a novel reinforcement learning method such that agents are evolved jointly towards successful communication and abstract graphical conventions. To inspect the emerged conventions, we define three key properties -- iconicity, symbolicity, and semanticity -- and design evaluation methods accordingly. Our experimental results under different controls are consistent with the observation in studies of human graphical conventions. Of note, we find that evolved sketches can preserve the continuum of semantics under proper environmental pressures. More interestingly, co-evolved agents can switch between conventionalized and iconic communication based on their familiarity with referents. We hope the present research can pave the path for studying emergent communication with the modality of sketches.

## [Falconn++: A Locality-sensitive Filtering Approach for Approximate Nearest Neighbor Search](#)

- Ninh Pham · Tao Liu
- abstract@[open-review](#): We present Falconn++, a novel locality-sensitive filtering (LSF) approach for approximate nearest neighbor search on angular distance. Falconn++ can filter out potential far away points in any hash bucket before querying, which results in higher quality candidates compared to other hashing-based solutions. Theoretically, Falconn++ asymptotically achieves lower query time complexity than Falconn, an optimal locality-sensitive hashing scheme on angular distance. Empirically, Falconn++ achieves a higher recall-speed tradeoff than Falconn on many real-world data sets. Falconn++ is also competitive against HNSW, an efficient representative of graph-based solutions on high search recall regimes.

## [Language Models with Image Descriptors are Strong Few-Shot Video-Language Learners](#)

- Zhenhai long Wang · Manling Li · Ruochen Xu · Luowei Zhou · Jie Lei · Xudong Lin · Shuohang Wang · Ziyi Yang · Chenguang Zhu · Derek Hoiem · Shih-Fu Chang · Mohit Bansal · Heng Ji
- abstract@[open-review](#): The goal of this work is to build flexible video-language models that can generalize to various video-to-text tasks from few examples. Existing few-shot video-language learners focus exclusively on the encoder, resulting in the absence of a video-to-text decoder to handle generative tasks. Video captioners have been pretrained on large-scale video-language datasets, but they rely heavily on finetuning and lack the ability to generate text for unseen tasks in a few-shot setting. We propose VidIL, a few-shot Video-language Learner via Image and Language models, which demonstrates strong performance on few-shot video-to-text tasks without the necessity of pretraining or finetuning on any video datasets. We use image-language models to translate the video content into frame captions, object, attribute, and event phrases, and compose them into a temporal-aware template. We then instruct a language model, with a prompt containing a few in-context examples, to generate a target output from the composed content. The

flexibility of prompting allows the model to capture any form of text input, such as automatic speech recognition (ASR) transcripts. Our experiments demonstrate the power of language models in understanding videos on a wide variety of video-language tasks, including video captioning, video question answering, video caption retrieval, and video future event prediction. Especially, on video future event prediction, our few-shot model significantly outperforms state-of-the-art supervised models trained on large-scale video datasets. Code and processed data are publicly available for research purposes at <https://github.com/MikeWangWZHL/VidIL>.

## [Efficient Frameworks for Generalized Low-Rank Matrix Bandit Problems](#)

- Yue Kang · Cho-Jui Hsieh · Thomas Chun Man Lee
- abstract@[open-review](#): In the stochastic contextual low-rank matrix bandit problem, the expected reward of an action is given by the inner product between the action's feature matrix and some fixed, but initially unknown  $\$d\_1\$$  by  $\$d\_2\$$  matrix  $\$\\Theta^*\$$  with rank  $\$r \\leq \\{d\_1, d\_2\\}\$$ , and an agent sequentially takes actions based on past experience to maximize the cumulative reward. In this paper, we study the generalized low-rank matrix bandit problem, which has been recently proposed in \cite{lu2021low} under the Generalized Linear Model (GLM) framework. To overcome the computational infeasibility and theoretical restraint of existing algorithms on this problem, we first propose the G-ESTT framework that modifies the idea from \cite{jun2019bilinear} by using Stein's method on the subspace estimation and then leverage the estimated subspaces via a regularization idea. Furthermore, we remarkably improve the efficiency of G-ESTT by using a novel exclusion idea on the estimated subspace instead, and propose the G-ESTS framework. We also show that both of our methods are the first algorithm to achieve the optimal  $\$\\tilde{O}\\((d\_1+d\_2)r\\sqrt{T})\$$  bound of regret presented in \cite{lu2021low} up to logarithm terms under some mild conditions, which improves upon the current regret of  $\$\\tilde{O}\\((d\_1+d\_2)^{3/2}r\\sqrt{rT})\$$ ~\cite{lu2021low}. For completeness, we conduct experiments to illustrate that our proposed algorithms, especially G-ESTS, are also computationally tractable and consistently outperform other state-of-the-art (generalized) linear matrix bandit methods based on a suite of simulations.

## [Knowledge Distillation: Bad Models Can Be Good Role Models](#)

- Gal Kaplun · Eran Malach · Preetum Nakkiran · Shai Shalev-Shwartz
- abstract@[open-review](#): Large neural networks trained in the overparameterized regime are able to fit noise to zero train error. Recent work of Nakkiran and Bansal has empirically observed that such networks behave as “conditional samplers” from the noisy distribution. That is, they replicate the noise in the train data to unseen examples. We give a theoretical framework for studying this conditional sampling behavior in the context of learning theory. We relate the notion of such samplers to knowledge distillation, where a student network imitates the outputs of a teacher on unlabeled data. We show that samplers, while being bad classifiers, can be good teachers. Concretely, we prove that distillation from samplers is guaranteed to produce a student which approximates the Bayes optimal classifier. Finally, we show that some common learning algorithms (e.g., Nearest-Neighbours and Kernel Machines) can often generate samplers when applied in the overparameterized regime.

## [Confidence-based Reliable Learning under Dual Noises](#)

- Peng Cui · Yang Yue · Zhipeng Deng · Jun Zhu
- abstract@[open-review](#): Deep neural networks (DNNs) have achieved remarkable success in a variety of computer vision tasks, where massive labeled images are routinely required for model optimization. Yet, the data collected from the open world are unavoidably polluted by noise, which may significantly undermine the efficacy of the learned models. Various attempts have been made to reliably train DNNs under data noise, but they separately account for either the noise existing in the labels or that existing in the images. A naive combination of the two lines of works would suffer from the limitations in both sides, and miss the opportunities to handle the two kinds of noise in parallel. This work provides a first, unified framework for reliable learning under the joint (image, label)-noise. Technically, we develop a confidence-based sample filter to progressively filter out noisy data without the need of pre-specifying noise ratio. Then, we penalize the model uncertainty of the detected noisy data instead of letting the model continue over-fitting the misleading information in them. Experiment results on various challenging synthetic and real-world noisy datasets verify that the proposed method can outperform competing baselines in the aspect of classification performance.

## [Spatial Mixture-of-Experts](#)

- Nikoli Dryden · Torsten Hoefer
- abstract@[open-review](#): Many data have an underlying dependence on spatial location; it may be weather on the Earth, a simulation on a mesh, or a registered image. Yet this feature is rarely taken advantage of, and violates common assumptions made by many neural network layers, such as translation equivariance. Further, many works that do incorporate locality fail to capture fine-grained structure. To address this, we introduce the Spatial Mixture-of-Experts (SMoE) layer, a sparsely-gated layer that learns spatial structure in the input domain and routes experts at a fine-grained level to utilize it. We also develop new techniques to train SMoEs, including a self-supervised routing loss and damping expert errors. Finally, we show strong results for SMoEs on numerous tasks, and set new state-of-the-art results for medium-range weather prediction.

## [EF-BV: A Unified Theory of Error Feedback and Variance Reduction Mechanisms for Biased and Unbiased Compression in Distributed Optimization](#)

- Laurent Condat · Kai Yi · Peter Richtarik
- abstract@[open-review](#): In distributed or federated optimization and learning, communication between the different computing units is often the bottleneck and gradient compression is widely used to reduce the number of bits sent within each communication round of iterative methods. There are two classes of compression operators and separate algorithms making use of them. In the case of unbiased random compressors with bounded variance (e.g., rand-k), the DIANA algorithm of Mishchenko et al. (2019), which implements a variance reduction technique for handling the variance introduced by compression, is the current state of the art. In the case of biased and contractive compressors (e.g., top-k), the EF21 algorithm of Richtárik et al. (2021), which instead implements an error-feedback mechanism, is the current state of the art. These two classes of compression schemes and algorithms are distinct, with different analyses and proof techniques. In this paper, we unify them into a single framework and propose a new algorithm, recovering DIANA and EF21 as particular cases. Our general approach works with a new, larger class of compressors, which has two parameters, the bias and the variance, and includes unbiased and biased compressors as particular cases. This allows us to inherit the best of the two worlds: like EF21 and unlike DIANA, biased compressors, like top-k, whose good performance in practice is recognized, can be used. And like DIANA and unlike EF21, independent randomness at the compressors allows to mitigate the effects of compression, with the convergence rate improving when the number of parallel workers is large. This is the first time that an algorithm with all these features is proposed. We prove its linear convergence under certain conditions. Our approach takes a step towards better understanding of two so-far distinct worlds of communication-efficient distributed learning.

## [Depth is More Powerful than Width in Deep Forest](#)

- Shen-Huan Lyu · Yi-Xiao He · Zhi-Hua Zhou
- abstract@[open-review](#): Random Forest (RF) is an ensemble learning algorithm proposed by \citet{breiman2001random} that constructs a large number of randomized decision trees individually and aggregates their predictions by naive averaging. \citet{zhou2019deep} further propose Deep Forest (DF) algorithm with multi-layer feature transformation, which significantly outperforms single-layer random forest in various application fields. Despite its great successes, little is known about the mathematical properties of the cascade structure. In this paper, we analyze the influence of depth and width on the consistency of cascade forests. Especially when the individual tree is inconsistent (as in practice, the individual tree is often set to be fully grown, i.e.,

there is only one sample at each leaf node), we find that the convergence rate of two-layer DF \textit{w.r.t.} the number of trees  $M$  can reach  $\mathcal{O}(1/M^2)$  under some mild conditions, while the convergence rate of single-layer RF is  $\mathcal{O}(1/M)$ . Therefore, learning decision trees in the "deep" layer will be more powerful than learning in the "shallow" layer. Experiments further confirm the theoretical advantages.

## [TREC: Transient Redundancy Elimination-based Convolution](#)

- Jiawei Guan Â· Feng Zhang Â· Jiesong Liu Â· Hsin-Hsuan Sung Â· Ruofan Wu Â· Xiaoyong Du Â· Xipeng Shen
- abstract@[open-review](#): The intensive computations in convolutional neural networks (CNNs) pose challenges for resource-constrained devices; eliminating redundant computations from convolution is essential. This paper gives a principled method to detect and avoid transient redundancy, a type of redundancy existing in input data or activation maps and hence changing across inferences. By introducing a new form of convolution (TREC), this new method makes transient redundancy detection and avoidance an inherent part of the CNN architecture, and the determination of the best configurations for redundancy elimination part of CNN backward propagation. We provide a rigorous proof of the robustness and convergence of TREC-equipped CNNs. TREC removes over 96% computations and achieves 3.51x average speedups on microcontrollers with minimal (about 0.7%) accuracy loss.

## [Local Linear Convergence of Gradient Methods for Subspace Optimization via Strict Complementarity](#)

- Dan Garber Â· Ron Fisher
- abstract@[open-review](#): We consider optimization problems in which the goal is find a  $k$ -dimensional subspace of  $\mathbb{R}^n$ ,  $k <$

## [Masked Generative Adversarial Networks are Robust Generation Learners](#)

- Jiaxing Huang Â· Kaiwen Cui Â· Dayan Guan Â· Aoran Xiao Â· Fangneng Zhan Â· Shijian Lu Â· Shengcai Liao Â· Eric Xing
- abstract@[open-review](#): This paper shows that masked generative adversarial network (MaskedGAN) is robust image generation learners with limited training data. The idea of MaskedGAN is simple: it randomly masks out certain image information for effective GAN training with limited data. We develop two masking strategies that work along orthogonal dimensions of training images, including a shifted spatial masking that masks the images in spatial dimensions with random shifts, and a balanced spectral masking that masks certain image spectral bands with self-adaptive probabilities. The two masking strategies complement each other which together encourage more challenging holistic learning from limited training data, ultimately suppressing trivial solutions and failures in GAN training. Albeit simple, extensive experiments show that MaskedGAN achieves superior performance consistently across different network architectures (e.g., CNNs including BigGAN and StyleGAN-v2 and Transformers including TransGAN and GANformer) and datasets (e.g., CIFAR-10, CIFAR-100, ImageNet, 100-shot, AFHQ, FFHQ and Cityscapes).

## [Asymptotically Unbiased Instance-wise Regularized Partial AUC Optimization: Theory and Algorithm](#)

- HuiYang Shao Â· Qianqian Xu Â· Zhiyong Yang Â· Shilong Bao Â· Qingming Huang
- abstract@[open-review](#): The Partial Area Under the ROC Curve (PAUC), typically including One-way Partial AUC (OPAUC) and Two-way Partial AUC (TPAUC), measures the average performance of a binary classifier within a specific false positive rate and/or true positive rate interval, which is a widely adopted measure when decision constraints must be considered. Consequently, PAUC optimization has naturally attracted increasing attention in the machine learning community within the last few years. Nonetheless, most of the existing methods could only optimize PAUC approximately, leading to inevitable biases that are not controllable. Fortunately, a recent work presents an unbiased formulation of the PAUC optimization problem via distributional robust optimization. However, it is based on the pair-wise formulation of AUC, which suffers from the limited scalability w.r.t. sample size and a slow convergence rate, especially for TPAUC. To address this issue, we present a simpler reformulation of the problem in an asymptotically unbiased and instance-wise manner. For both OPAUC and TPAUC, we come to a nonconvex strongly concave min-max regularized problem of instance-wise functions. On top of this, we employ an efficient solver that enjoys a linear per-iteration computational complexity w.r.t. the sample size and a time-complexity of  $\mathcal{O}(\epsilon^{-1/3})$  to reach a  $\epsilon$  stationary point. Furthermore, we find that the min-max reformulation also facilitates the theoretical analysis of generalization error as a byproduct. Compared with the existing results, we present new error bounds that are much easier to prove and could deal with hypotheses with real-valued outputs. Finally, extensive experiments on several benchmark datasets demonstrate the effectiveness of our method.

## [Non-stationary Transformers: Rethinking the Stationarity in Time Series Forecasting](#)

- Yong Liu Â· Haixu Wu Â· Jianmin Wang Â· Mingsheng Long
- abstract@[open-review](#): Transformers have shown great power in time series forecasting due to their global-range modeling ability. However, their performance can degenerate terribly on non-stationary real-world data in which the joint distribution changes over time. Previous studies primarily adopt stationarization to reduce the non-stationarity of original series for better predictability. But the stationarized series deprived of inherent non-stationarity can be less instructive for real-world bursty events forecasting. This problem, termed over-stationarization in this paper, leads Transformers to generate indistinguishable temporal attentions for different series and impedes the predictive capability of deep models. To tackle the dilemma between series predictability and model capability, we propose Non-stationary Transformers as a generic framework with two interdependent modules: Series Stationarization and De-stationary Attention. Concretely, Series Stationarization unifies the statistics of each input and converts the output with restored statistics for better predictability. To address over-stationarization, De-stationary Attention is devised to recover the intrinsic non-stationary information into temporal dependencies by approximating distinguishable attentions learned from unstationarized series. Our Non-stationary Transformers framework consistently boosts mainstream Transformers by a large margin, which reduces 49.43% MSE on Transformer, 47.34% on Informer, and 46.89% on Reformer, making them the state-of-the-art in time series forecasting.

## [Fully Convolutional One-Stage 3D Object Detection on LiDAR Range Images](#)

- Zhi Tian Â· Xiangxiang Chu Â· Xiaoming Wang Â· Xiaolin Wei Â· Chunhua Shen
- abstract@[open-review](#): We present a simple yet effective fully convolutional one-stage 3D object detector for LiDAR point clouds of autonomous driving scenes, termed FCOS-LiDAR. Unlike the dominant methods that use the bird-eye view (BEV), our proposed detector detects objects from the range view (RV, a.k.a. range image) of the LiDAR points. Due to the range view's compactness and compatibility with the LiDAR sensors' sampling process on self-driving cars, the range view-based object detector can be realized by solely exploiting the vanilla 2D convolutions, departing from the BEV-based methods which often involve complicated voxelization operations and sparse convolutions. For the first time, we show that an RV-based 3D detector with standard 2D convolutions alone can achieve comparable performance to state-of-the-art BEV-based detectors while being significantly faster and simpler. More importantly, almost all previous range view-based detectors only focus on single-frame point clouds since it is challenging to fuse multi-frame point clouds into a single range view. In this work, we tackle this challenging issue with a novel range view projection mechanism, and for the first time demonstrate the benefits of fusing multi-frame point clouds for a range-view based detector. Extensive experiments on nuScenes show the superiority of our proposed method and we believe that our work can be strong evidence that an RV-based 3D detector can compare favourably with the current mainstream BEV-based detectors.

## [Counterfactual Fairness with Partially Known Causal Graph](#)

- Aoqi Zuo · Susan Wei · Tongliang Liu · Bo Han · Kun Zhang · Mingming Gong
- abstract@[open-review](#): Fair machine learning aims to avoid treating individuals or sub-populations unfavourably based on sensitive attributes, such as gender and race. Those methods in fair machine learning that are built on causal inference ascertain discrimination and bias through causal effects. Though causality-based fair learning is attracting increasing attention, current methods assume the true causal graph is fully known. This paper proposes a general method to achieve the notion of counterfactual fairness when the true causal graph is unknown. To be able to select features that lead to counterfactual fairness, we derive the conditions and algorithms to identify ancestral relations between variables on a Partially Directed Acyclic Graph (PDAG), specifically, a class of causal DAGs that can be learned from observational data combined with domain knowledge. Interestingly, we find that counterfactual fairness can be achieved as if the true causal graph were fully known, when specific background knowledge is provided: the sensitive attributes do not have ancestors in the causal graph. Results on both simulated and real-world datasets demonstrate the effectiveness of our method.

## [How Sampling Impacts the Robustness of Stochastic Neural Networks](#)

- Sina Dabrener · Asja Fischer
- abstract@[open-review](#): Stochastic neural networks (SNNs) are random functions whose predictions are gained by averaging over multiple realizations. Consequently, a gradient-based adversarial example is calculated based on one set of samples and its classification on another set. In this paper we derive a sufficient condition for such a stochastic prediction to be robust against a given sample-based attack. This allows us to identify the factors that lead to an increased robustness of SNNs and gives theoretical explanations for: (i) the well known observation, that increasing the amount of samples drawn for the estimation of adversarial examples increases the attack's strength,(ii) why increasing the number of samples during an attack can not fully reduce the effect of stochasticity (iii) why the sample size during inference does not influence the robustness, and(iv) why a higher gradient variance and shorter expected value of the gradient relates to a higher robustness. Our theoretical findings give a unified view on the mechanisms underlying previously proposed approaches for increasing attack strengths or model robustness, which we verify by an extensive empirical analysis.

## [Monte Carlo Augmented Actor-Critic for Sparse Reward Deep Reinforcement Learning from Suboptimal Demonstrations](#)

- Albert Wilcox · Ashwin Balakrishna · Daniel Brown · Jules Dedieu · Wyame Benslimane · Ken Goldberg
- abstract@[open-review](#): Providing densely shaped reward functions for RL algorithms is often exceedingly challenging, motivating the development of RL algorithms that can learn from easier-to-specify sparse reward functions. This sparsity poses new exploration challenges; one common response is to use demonstrations to provide initial signal about regions of the state space with high rewards. However, prior RL from demonstrations algorithms introduce significant complexity and many hyperparameters, making them hard to implement and tune. We introduce Monte Carlo Augmented Actor-Critic (MCAC), a parameter free modification to standard actor-critic algorithms which initializes the replay buffer with demonstrations and computes a modified Q-value by taking the maximum of the standard temporal distance (TD) target and a Monte Carlo estimate of the reward-to-go. This encourages exploration in the neighborhood of high-performing trajectories by encouraging high Q-values in corresponding regions of the state space. Experiments across 5 continuous control domains suggest that MCAC can be used to significantly increase learning efficiency for a number of prior RL and RL-from-demonstrations algorithms.

## [Robustness in deep learning: The width \(good\), the depth \(bad\), and the initialization \(ugly\)](#)

- Zhenyu Zhu · Fanghui Liu · Grigoris Chrysos · Volkan Cevher
- abstract@[open-review](#): We study the average robustness notion in deep neural networks in (selected) wide and narrow, deep and shallow, as well as lazy and non-lazy training settings. We prove that in the under-parameterized setting, width has a negative effect while it improves robustness in the over-parameterized setting. The effect of depth closely depends on the initialization and the training mode. In particular, when initialized with LeCun initialization, depth helps robustness with lazy training regime. In contrast, when initialized with Neural Tangent Kernel (NTK) and He-initialization, depth hurts the robustness. Moreover, under non-lazy training regime, we demonstrate how the width of a two-layer ReLU network benefits robustness. Our theoretical developments improve the results by [Huang et al. NeurIPS21; Wu et al. NeurIPS21] and are consistent with [Bubeck and Sellke NeurIPS21; Bubeck et al. COLT21].

## [Animatable 3D-Aware Face Image Generation for Realistic Video Avatars](#)

- Yue Wu · Yu Deng · Jiaolong Yang · Fangyun Wei · Qifeng Chen · Xin Tong
- abstract@[open-review](#): Face image generation and animation have been a longstanding task. Although many 2D generative models yield excellent manipulations in 2D space, they often suffer from 3D inconsistency and undesirable artifacts when rendering from different camera viewpoints, and thus are not suitable for animations in video. Recently, 3D-aware GANs extend 2D GANs by using underlying 3D representations. Although these methods can preserve the 3D consistency across different viewpoints, they cannot achieve fine-grained control over attributes, most importantly, facial expression. In this paper, we propose an animatable 3D-aware face image generation method. Our framework mainly consists of a template implicit field and a 3D deformation field. The template field represents the canonical space and is shared across the same identity. Different expressions can be generated by deforming the manifolds in the target space to the canonical space. We enforce the generation to follow a prior 3D face parametric model by incorporating 3D-level imitative learning to encourage the deformation field to follow 3D priors. Experiments show our method can produce high-quality animatable video avatars with strong visual 3D consistency.

## [Learning with little mixing](#)

- Ingvar Ziemann · Stephen Tu
- abstract@[open-review](#): We study square loss in a realizable time-series framework with martingale difference noise. Our main result is a fast rate excess risk bound which shows that whenever a trajectory hypercontractivity condition holds, the risk of the least-squares estimator on dependent data matches the iid rate order-wise after a burn-in time. In comparison, many existing results in learning from dependent data have rates where the effective sample size is deflated by a factor of the mixing-time of the underlying process, even after the burn-in time. Furthermore, our results allow the covariate process to exhibit long range correlations which are substantially weaker than geometric ergodicity. We call this phenomenon learning with little mixing, and present several examples for when it occurs: bounded function classes for which the  $L^2$  and  $L^{2+\epsilon}$  norms are equivalent, finite state irreducible and aperiodic Markov chains, various parametric models, and a broad family of infinite dimensional  $\ell^2(\mathbb{N})$  ellipsoids. By instantiating our main result to system identification of nonlinear dynamics with generalized linear model transitions, we obtain a nearly minimax optimal excess risk bound after only a polynomial burn-in time.

## [Globally Optimal Algorithms for Fixed-Budget Best Arm Identification](#)

- Junpei Komiyama · Taira Tsuchiya · Junya Honda
- abstract@[open-review](#): We consider the fixed-budget best arm identification problem where the goal is to find the arm of the largest mean with a fixed number of samples. It is known that the probability of misidentifying the best arm is exponentially small to the number of rounds. However, limited characterizations have been discussed on the rate (exponent) of this value. In this paper, we characterize the optimal rate as a result of global optimization over all possible parameters. We introduce two rates,  $R^{\mathcal{M}(\mathcal{G})}$  and  $R^{\mathcal{M}(\mathcal{G})^{infty}}$ , corresponding to lower bounds on the misidentification probability, each of which is associated with a proposed algorithm. The rate  $R^{\mathcal{M}(\mathcal{G})}$  is associated with  $R^{\mathcal{M}(\mathcal{G})}$ -tracking, which can be efficiently implemented by a neural network and is shown to outperform existing algorithms. However, this rate requires a

*nontrivial condition to be achievable. To deal with this issue, we introduce the second rate  $R^{\mathcal{G}} \infty$ . We show that this rate is indeed achievable by introducing a conceptual algorithm called delayed optimal tracking (DOT).*

## [Incorporating Bias-aware Margins into Contrastive Loss for Collaborative Filtering](#)

- An Zhang · Wenchang Ma · Xiang Wang · Tat-Seng Chua
- abstract@[open-review](#): Collaborative filtering (CF) models easily suffer from popularity bias, which makes recommendation deviate from users' actual preferences. However, most current debiasing strategies are prone to playing a trade-off game between head and tail performance, thus inevitably degrading the overall recommendation accuracy. To reduce the negative impact of popularity bias on CF models, we incorporate Bias-aware margins into Contrastive loss and propose a simple yet effective BC Loss, where the margin tailors quantitatively to the bias degree of each user-item interaction. We investigate the geometric interpretation of BC loss, then further visualize and theoretically prove that it simultaneously learns better head and tail representations by encouraging the compactness of similar users/items and enlarging the dispersion of dissimilar users/items. Over six benchmark datasets, we use BC loss to optimize two high-performing CF models. In various evaluation settings (i.e., imbalanced/balanced, temporal split, fully-observed unbiased, tail/head test evaluations), BC loss outperforms the state-of-the-art debiasing and non-debiasing methods with remarkable improvements. Considering the theoretical guarantee and empirical success of BC loss, we advocate using it not just as a debiasing strategy, but also as a standard loss in recommender models. Codes are available at <https://anonymous.4open.science/r/BC-Loss-9DF2>.

## [Hierarchical Graph Transformer with Adaptive Node Sampling](#)

- ZAIKI ZHANG · Qi Liu · Qingyong Hu · Chee-Kong Lee
- abstract@[open-review](#): The Transformer architecture has achieved remarkable success in a number of domains including natural language processing and computer vision. However, when it comes to graph-structured data, transformers have not achieved competitive performance, especially on large graphs. In this paper, we identify the main deficiencies of current graph transformers: (1) Existing node sampling strategies in Graph Transformers are agnostic to the graph characteristics and the training process. (2) Most sampling strategies only focus on local neighbors and neglect the long-range dependencies in the graph. We conduct experimental investigations on synthetic datasets to show that existing sampling strategies are sub-optimal. To tackle the aforementioned problems, we formulate the optimization strategies of node sampling in Graph Transformer as an adversary bandit problem, where the rewards are related to the attention weights and can vary in the training procedure. Meanwhile, we propose a hierarchical attention scheme with graph coarsening to capture the long-range interactions while reducing computational complexity. Finally, we conduct extensive experiments on real-world datasets to demonstrate the superiority of our method over existing graph transformers and popular GNNs.

## [Where to Pay Attention in Sparse Training for Feature Selection?](#)

- Ghada Sokar · Zahra Atashgahi · Mykola Pechenizkiy · Decebal Constantin Mocanu
- abstract@[open-review](#): A new line of research for feature selection based on neural networks has recently emerged. Despite its superiority to classical methods, it requires many training iterations to converge and detect the informative features. For datasets with a large number of samples or a very high dimensional feature space, the computational time becomes prohibitively long. In this paper, we present a new efficient unsupervised method for feature selection based on sparse autoencoders. In particular, we propose a new sparse training algorithm that optimizes a model's sparse topology during training to quickly pay attention to informative features. The attention-based adaptation of the sparse topology enables fast detection of informative features after a few training iterations. We performed extensive experiments on 10 datasets of different types, including image, speech, text, artificial, and biological. They cover a wide range of characteristics, such as low and high-dimensional feature spaces, as well as few and large training samples. Our proposed approach outperforms the state-of-the-art methods in terms of the selection of informative features while reducing training iterations and computational costs substantially. Moreover, the experiments show the robustness of our method in extremely noisy environments.

## [Mutual Information Divergence: A Unified Metric for Multimodal Generative Models](#)

- Jin-Hwa Kim · Yunji Kim · Jiyoung Lee · Kang Min Yoo · Sang-Woo Lee
- abstract@[open-review](#): Text-to-image generation and image captioning are recently emerged as a new experimental paradigm to assess machine intelligence. They predict continuous quantity accompanied by their sampling techniques in the generation, making evaluation complicated and intractable to get marginal distributions. Based on a recent trend that multimodal generative evaluations exploit a vision-and-language pre-trained model, we propose the negative Gaussian cross-mutual information using the CLIP features as a unified metric, coined by Mutual Information Divergence (MID). To validate, we extensively compare it with competing metrics using carefully-generated or human-annotated judgments in text-to-image generation and image captioning tasks. The proposed MID significantly outperforms the competitive methods by having consistency across benchmarks, sample parsimony, and robustness toward the exploited CLIP model. We look forward to seeing the underrepresented implications of the Gaussian cross-mutual information in multimodal representation learning and the future works based on this novel proposition.

## [Outlier Suppression: Pushing the Limit of Low-bit Transformer Language Models](#)

- Xiuying Wei · Yunchen Zhang · Xiangguo Zhang · Ruihao Gong · Shanghang Zhang · Qi Zhang · Fengwei Yu · Xianglong Liu
- abstract@[open-review](#): Transformer architecture has become the fundamental element of the widespread natural language processing~(NLP) models. With the trends of large NLP models, the increasing memory and computation costs hinder their efficient deployment on resource-limited devices. Therefore, transformer quantization attracts wide research interest. Recent work recognizes that structured outliers are the critical bottleneck for quantization performance. However, their proposed methods increase the computation overhead and still leave the outliers there. To fundamentally address this problem, this paper delves into the inherent inducement and importance of the outliers. We discover that  $\boldsymbol{\gamma}$  in LayerNorm (LN) acts as a sinful amplifier for the outliers, and the importance of outliers varies greatly where some outliers provided by a few tokens cover a large area but can be clipped sharply without negative impacts. Motivated by these findings, we propose an outlier suppression framework including two components: Gamma Migration and Token-Wise Clipping. The Gamma Migration migrates the outlier amplifier to subsequent modules in an equivalent transformation, contributing to a more quantization-friendly model without any extra burden. The Token-Wise Clipping takes advantage of the large variance of token range and designs a token-wise coarse-to-fine pipeline, obtaining a clipping range with minimal final quantization loss in an efficient way. This framework effectively suppresses the outliers and can be used in a plug-and-play mode. Extensive experiments prove that our framework surpasses the existing works and, for the first time, pushes the 6-bit post-training BERT quantization to the full-precision (FP) level. Our code is available at [https://github.com/wimh966/outlier\\_suppression](https://github.com/wimh966/outlier_suppression).

## [Self-Supervised Multi-Granularity Map Learning for Vision-and-Language Navigation](#)

- Peihao Chen · Dongyu Ji · Kunyang Lin · Runhao Zeng · Thomas Li · Mingkui Tan · Chuang Gan
- abstract@[open-review](#): We address a practical yet challenging problem of training robot agents to navigate in an environment following a path described by some language instructions. The instructions often contain descriptions of objects in the environment and path cues defined by humans. To achieve accurate and efficient navigation, it is critical to build a map that accurately represents both spatial location and the semantic information of the environment objects. However, enabling a robot to build a map that well represents the environment is extremely challenging as the environment often involves diverse objects with various attributes. In this paper, we propose a multi-granularity map, which contains both object fine-grained details (eg, color, texture) and semantic classes, to represent objects more comprehensively. Moreover, we propose a weakly-supervised auxiliary task, which requires the agent to localize instruction-relevant objects on the map. Through this task, the agent not only learns to localize the instruction-relevant objects for

navigation but also is encouraged to learn a better map representation that reveals object information. We then feed the learned map and instruction to a waypoint predictor to determine the next navigation goal. Experimental results show our method outperforms the state-of-the-art by 4.0% and 4.6% w.r.t. success rate both in seen and unseen environments, respectively on VLN-CE dataset.

## [Two-Stream Network for Sign Language Recognition and Translation](#)

- Yutong Chen · Ronglai Zuo · Fangyun Wei · Yu Wu · Shujie LIU · Brian Mak
- abstract@[open-review](#): Sign languages are visual languages using manual articulations and non-manual elements to convey information. For sign language recognition and translation, the majority of existing approaches directly encode RGB videos into hidden representations. RGB videos, however, are raw signals with substantial visual redundancy, leading the encoder to overlook the key information for sign language understanding. To learn more meaningful representations and incorporate domain knowledge, such as handshape and facial expressions, we introduce a dual visual encoder containing two separate streams to model both the raw videos and the keypoint sequences generated by an off-the-shelf keypoint estimator. To make the two streams interact with each other, we explore a variety of techniques, including bidirectional lateral connection, sign pyramid network with auxiliary supervision, and frame-level self-distillation. The resulting model is called TwoStream-SLR, which is competent for sign language recognition (SLR). TwoStream-SLR is extended to a sign language translation (SLT) model, TwoStream-SLT, by simply attaching an extra translation network. Experimentally, our TwoStream-SLR and TwoStream-SLT achieve state-of-the-art performance on SLR and SLT tasks across a series of datasets including Phoenix-2014, Phoenix-2014T, and CSL-Daily.

## [TPU-KNN: K Nearest Neighbor Search at Peak FLOP/s](#)

- Felix Chern · Blake Hechtman · Andy Davis · Ruiqi Guo · David Majnemer · Sanjiv Kumar
- abstract@[open-review](#): This paper presents a novel nearest neighbor search algorithm achieving TPU (Google Tensor Processing Unit) peak performance, outperforming state-of-the-art GPU algorithms with similar level of recall. The design of the proposed algorithm is motivated by an accurate accelerator performance model that takes into account both the memory and instruction bottlenecks. Our algorithm comes with an analytical guarantee of recall in expectation and does not require maintaining sophisticated index data structure or tuning, making it suitable for applications with frequent updates. Our work is available in the open-source package of Jax and Tensorflow on TPU.

## [Bootstrapped Transformer for Offline Reinforcement Learning](#)

- Kerong Wang · Hanye Zhao · Xufang Luo · Kan Ren · Weinan Zhang · Dongsheng Li
- abstract@[open-review](#): Offline reinforcement learning (RL) aims at learning policies from previously collected static trajectory data without interacting with the real environment. Recent works provide a novel perspective by viewing offline RL as a generic sequence generation problem, adopting sequence models such as Transformer architecture to model distributions over trajectories and repurposing beam search as a planning algorithm. However, the training datasets utilized in general offline RL tasks are quite limited and often suffering from insufficient distribution coverage, which could be harmful to training sequence generation models yet has not drawn enough attention in the previous works. In this paper, we propose a novel algorithm named Bootstrapped Transformer, which incorporates the idea of bootstrapping and leverages the learned model to self-generate more offline data to further boost the training of sequence model. We conduct extensive experiments on two offline RL benchmarks and demonstrate that our model can largely remedy the limitations of the existing offline RL training and beat other strong baseline methods. We also analyze the generated pseudo data and the revealed characteristics may shed some light on offline RL training.

## [Gradient-Free Methods for Deterministic and Stochastic Nonsmooth Nonconvex Optimization](#)

- Tianyi Lin · Zeyu Zheng · Michael Jordan
- abstract@[open-review](#): Nonsmooth nonconvex optimization problems broadly emerge in machine learning and business decision making, whereas two core challenges impede the development of efficient solution methods with finite-time convergence guarantee: the lack of computationally tractable optimality criterion and the lack of computationally powerful oracles. The contributions of this paper are two-fold. First, we establish the relationship between the celebrated Goldstein subdifferential~\citep{Goldstein-1977-Optimization} and uniform smoothing, thereby providing the basis and intuition for the design of gradient-free methods that guarantee the finite-time convergence to a set of Goldstein stationary points. Second, we propose the gradient-free method (GFM) and stochastic GFM for solving a class of nonsmooth nonconvex optimization problems and prove that both of them can return a  $\$ (\delta, \epsilon) \$$ -Goldstein stationary point of a Lipschitz function  $f$  at an expected convergence rate at  $\$ O(d^{3/2} \delta^{-1} \epsilon^{-4}) \$$  where  $d$  is the problem dimension. Two-phase versions of GFM and SGFM are also proposed and proven to achieve improved large-deviation results. Finally, we demonstrate the effectiveness of 2-SGFM on training ReLU neural networks with the \textsc{Minst} dataset.

## [Revisiting Optimal Convergence Rate for Smooth and Non-convex Stochastic Decentralized Optimization](#)

- Kun Yuan · Xinmeng Huang · Yiming Chen · Xiaohan Zhang · Yingya Zhang · PAN PAN
- abstract@[open-review](#): While numerous effective decentralized algorithms have been proposed with theoretical guarantees and empirical successes, the performance limits in decentralized optimization, especially the influence of network topology and its associated weight matrix on the optimal convergence rate, have not been fully understood. While Lu and Sa have recently provided an optimal rate for non-convex stochastic decentralized optimization using weight matrices associated with linear graphs, the optimal rate with general weight matrices remains unclear. This paper revisits non-convex stochastic decentralized optimization and establishes an optimal convergence rate with general weight matrices. In addition, we also establish the first optimal rate when non-convex loss functions further satisfy the Polyak-Lojasiewicz (PL) condition. Following existing lines of analysis in literature cannot achieve these results. Instead, we leverage the Ring-Lattice graph to admit general weight matrices while maintaining the optimal relation between the graph diameter and weight matrix connectivity. Lastly, we develop a new decentralized algorithm to attain the above two optimal rates up to logarithm factors.

## [Communication-efficient distributed eigenspace estimation with arbitrary node failures](#)

- Vasileios Charisopoulos · Anil Damle
- abstract@[open-review](#): We develop an eigenspace estimation algorithm for distributed environments with arbitrary node failures, where a subset of computing nodes can return structurally valid but otherwise arbitrarily chosen responses. Notably, this setting encompasses several important scenarios that arise in distributed computing and data-collection environments such as silent/soft errors, outliers or corrupted data at certain nodes, and adversarial responses. Our estimator builds upon and matches the performance of a recently proposed non-robust estimator up to an additive  $\$ \tilde{O}(\sigma \sqrt{\alpha}) \$$  error, where  $\sigma^2$  is the variance of the existing estimator and  $\alpha$  is the fraction of corrupted nodes.

## [Hierarchical classification at multiple operating points](#)

- Jack Valmadre
- abstract@[open-review](#): Many classification problems consider classes that form a hierarchy. Classifiers that are aware of this hierarchy may be able to make confident predictions at a coarse level despite being uncertain at the fine-grained level. While it is generally possible to vary the granularity of predictions using a threshold at inference time, most contemporary work considers only leaf-node prediction, and almost no prior work has compared

methods at multiple operating points. We present an efficient algorithm to produce operating characteristic curves for any method that assigns a score to every class in the hierarchy. Applying this technique to evaluate existing methods reveals that top-down classifiers are dominated by a naive flat softmax classifier across the entire operating range. We further propose two novel loss functions and show that a soft variant of the structured hinge loss is able to significantly outperform the flat baseline. Finally, we investigate the poor accuracy of top-down classifiers and demonstrate that they perform relatively well on unseen classes.

## [Noise Attention Learning](#)

- Yangdi Lu · Yang Bo · Wenbo He
- abstract@[open-review](#): Machine learning has been highly successful in data-driven applications but is often hampered when the data contains noise, especially label noise. When trained on noisy labels, deep neural networks tend to fit all noisy labels, resulting in poor generalization. To handle this problem, a common idea is to force the model to fit only clean samples rather than the mislabeled ones. In this paper, we propose a simple yet effective method that automatically distinguishes the mislabeled samples and prevents the model from memorizing them, named Noise Attention Learning. In our method, we introduce an attention branch to produce attention weights based on representations of samples. The attention branch is learned to divide the samples according to the predictive power in their representations. We design the corresponding loss function that incorporates the attention weights for training the model without affecting the original learning direction. Empirical results show that most of the mislabeled samples yield significantly lower weights than clean ones. Furthermore, our theoretical analysis shows that the gradients of training samples are dynamically scaled by the attention weights, implicitly preventing memorization of the mislabeled samples. Experimental results on two benchmarks (CIFAR-10 and CIFAR-100) and three real-world datasets (ANIMAL-10N, Clothing1M and Webvision) demonstrate that our approach outperforms state-of-the-art methods.

## [PopArt: Efficient Sparse Regression and Experimental Design for Optimal Sparse Linear Bandits](#)

- Kyoungseok Jang · Chicheng Zhang · Kwang-Sung Jun
- abstract@[open-review](#): In sparse linear bandits, a learning agent sequentially selects an action from a fixed action set and receives reward feedback, and the reward function depends linearly on a few coordinates of the covariates of the actions. This has applications in many real-world sequential decision making problems. In this paper, we devise a simple, novel sparse linear estimation method called  $\text{PopArt}$  that enjoys a tighter  $\ell_1$  recovery guarantee compared to Lasso (Tibshirani, 1996). Our bound naturally motivates an experimental design criterion that is convex and thus computationally efficient to solve. Based on our novel estimator and design criterion, we derive sparse linear bandit algorithms that enjoy improved regret upper bounds upon the state of the art (Hao et al., 2020), especially in terms of the geometry of the given action set. Finally, we prove a matching lower bound for sparse linear bandits in the data-poor regime, which closes the gap between upper and lower bounds in prior work.

## [Deep Equilibrium Approaches to Diffusion Models](#)

- Ashwini Pokle · Zhengyang Geng · J. Zico Kolter
- abstract@[open-review](#): Diffusion-based generative models have shown to be extremely effective in generating high-quality images, with generated samples often surpassing the quality of those produced by other models under several metrics. One distinguishing feature of these models, however, is that they typically require long sampling chains in order to produce high-fidelity images. This presents a challenge not only from the lenses of sampling time, but also from the inherent difficulty in backpropagating through these chains in order to accomplish tasks such as model inversion, i.e., approximately finding latent states that generate known images. In this paper, we look at diffusion models through a different perspective, that of a (deep) equilibrium (DEQ) fixed point model. Specifically, we extend the recent denoising diffusion implicit model (DDIM), and model the entire sampling chain as a joint, multi-variate fixed point system. This setup provides an elegant unification of diffusion and equilibrium models, and shows benefits in 1) single-shot image sampling, as it replaces the fully-serial typical sampling process with a parallel one; and 2) model inversion, where we can leverage fast gradients in the DEQ setting to much more quickly find the noise that generates a given image. The approach is also orthogonal and thus complementary to other methods used to reduce the sampling time, or improve model inversion. We demonstrate our method's strong performance across several datasets, including CIFAR10, CelebA, and LSUN Bedroom and Churches.

## [Improved Regret Analysis for Variance-Adaptive Linear Bandits and Horizon-Free Linear Mixture MDPs](#)

- Yeoneung Kim · Insoon Yang · Kwang-Sung Jun
- abstract@[open-review](#): In online learning problems, exploiting low variance plays an important role in obtaining tight performance guarantees yet is challenging because variances are often not known a priori. Recently, considerable progress has been made by Zhang et al. (2021) where they obtain a variance-adaptive regret bound for linear bandits without knowledge of the variances and a horizon-free regret bound for linear mixture Markov decision processes (MDPs). In this paper, we present novel analyses that improve their regret bounds significantly. For linear bandits, we achieve  $\tilde{O}(d^{1.5}\sqrt{\sum_{k=1}^K \sigma_k^2} + d^2)$  where  $d$  is the dimension of the features,  $K$  is the time horizon, and  $\sigma_k^2$  is the noise variance at time step  $k$ , and  $\tilde{O}$  ignores polylogarithmic dependence, which is a factor of  $d^3$  improvement. For linear mixture MDPs with the assumption of maximum cumulative reward in an episode being in  $[0,1]$ , we achieve a horizon-free regret bound of  $\tilde{O}(d\sqrt{K} + d^2)$  where  $d$  is the number of base models and  $K$  is the number of episodes. This is a factor of  $d^{3.5}$  improvement in the leading term and  $d^7$  in the lower order term. Our analysis critically relies on a novel peeling-based regret analysis that leverages the elliptical potential 'count' lemma.

## [Signal Propagation in Transformers: Theoretical Perspectives and the Role of Rank Collapse](#)

- Sotiris Anagnostidis · Luca Biggio · Lorenzo Noci · Antonio Orvieto · Sidak Pal Singh · Aurelien Lucchi
- abstract@[open-review](#): Transformers have achieved remarkable success in several domains, ranging from natural language processing to computer vision. Nevertheless, it has been recently shown that stacking self-attention layers – the distinctive architectural component of Transformers – can result in rank collapse of the tokens<sup>TM</sup> representations at initialization. The question of if and how rank collapse affects training is still largely unanswered, and its investigation is necessary for a more comprehensive understanding of this architecture. In this work, we shed new light on the causes and the effects of this phenomenon. First, we show that rank collapse of the tokens<sup>TM</sup> representations hinders training by causing the gradients of the queries and keys to vanish at initialization. Furthermore, we provide a thorough description of the origin of rank collapse and discuss how to prevent it via an appropriate depth-dependent scaling of the residual branches. Finally, our analysis unveils that specific architectural hyperparameters affect the gradients of queries, keys and values differently, leading to disproportionate gradient norms. This suggests an explanation for the widespread use of adaptive methods for Transformers' optimization.

## [Exploring the Latent Space of Autoencoders with Interventional Assays](#)

- Felix Leeb · Stefan Bauer · Michel Besserve · Bernhard Schölkopf
- abstract@[open-review](#): Autoencoders exhibit impressive abilities to embed the data manifold into a low-dimensional latent space, making them a staple of representation learning methods. However, without explicit supervision, which is often unavailable, the representation is usually uninterpretable, making analysis and principled progress challenging. We propose a framework, called latent responses, which exploits the locally contractive behavior exhibited by variational autoencoders to explore the learned manifold. More specifically, we develop tools to probe the representation using interventions in the latent space to quantify the relationships between latent variables. We extend the notion of disentanglement to take the learned generative process into account and consequently avoid the limitations of existing metrics that may rely on spurious correlations. Our analyses underscore the importance of

studying the causal structure of the representation to improve performance on downstream tasks such as generation, interpolation, and inference of the factors of variation.

## [Embodied Scene-aware Human Pose Estimation](#)

- Zhengyi Luo · Shun Iwase · Ye Yuan · Kris Kitani
- abstract@[open-review](#): We propose embodied scene-aware human pose estimation where we estimate 3D poses based on a simulated agent's proprioception and scene awareness, along with external third-person observations. Unlike prior methods that often resort to multistage optimization, non-causal inference, and complex contact modeling to estimate human pose and human scene interactions, our method is one stage, causal, and recovers global 3D human poses in a simulated environment. Since 2D third-person observations are coupled with the camera pose, we propose to disentangle the camera pose and use a multi-step projection gradient defined in the global coordinate frame as the movement cue for our embodied agent. Leveraging a physics simulation and prescanned scenes 3D mesh, we simulate our agent in everyday environments (libraries, offices, bedrooms, etc.) and equip our agent with environmental sensors to intelligently navigate and interact with scene geometries. Our method also relies only on 2D keypoints and can be trained on synthetic datasets derived from popular human motion databases. To evaluate, we use the popular H36M and PROX datasets and achieve high-quality pose estimation on the challenging PROX dataset without ever using PROX motion sequences for training.

## [Multi-block-Single-probe Variance Reduced Estimator for Coupled Compositional Optimization](#)

- Wei Jiang · Gang Li · Yibo Wang · Lijun Zhang · Tianbao Yang
- abstract@[open-review](#): Variance reduction techniques such as SPIDER/SARAH/STORM have been extensively studied to improve the convergence rates of stochastic non-convex optimization, which usually maintain and update a sequence of estimators for a single function across iterations. What if we need to track multiple functional mappings across iterations but only with access to stochastic samples of  $\mathcal{O}(1)$  functional mappings at each iteration? There is an important application in solving an emerging family of coupled compositional optimization problems in the form of  $\sum_{i=1}^m f_i(g_i(\mathbf{w}))$ , where  $g_i$  is accessible through a stochastic oracle. The key issue is to track and estimate a sequence of  $\mathbf{g}$  across iterations, where  $\mathbf{g} = (g_1(\mathbf{w}), \dots, g_m(\mathbf{w}))$  has  $m$  blocks and it is only allowed to probe  $\mathcal{O}(1)$  blocks to attain their stochastic values and Jacobians. To improve the complexity for solving these problems, we propose a novel stochastic method named Multi-block-Single-probe Variance Reduced (MSVR) estimator to track the sequence of  $\mathbf{g}$ . It is inspired by STORM but introduces a customized error correction term to alleviate the noise not only in stochastic samples for the selected blocks but also in those blocks that are not sampled. With the help of the MSVR estimator, we develop several algorithms for solving the aforementioned compositional problems with improved complexities across a spectrum of settings with non-convex/convex/strongly convex objectives. Our results improve upon prior ones in several aspects, including the order of sample complexities and dependence on the strong convexity parameter. Empirical studies on multi-task deep AUC maximization demonstrate the better performance of using the new estimator.

## [Regret Bounds for Risk-Sensitive Reinforcement Learning](#)

- Osbert Bastani · Jason Yecheng Ma · Estelle Shen · Wanqiao Xu
- abstract@[open-review](#): In safety-critical applications of reinforcement learning such as healthcare and robotics, it is often desirable to optimize risk-sensitive objectives that account for tail outcomes rather than expected reward. We prove the first regret bounds for reinforcement learning under a general class of risk-sensitive objectives including the popular CVaR objective. Our theory is based on a novel characterization of the CVaR objective as well as a novel optimistic MDP construction.

## [Wasserstein Logistic Regression with Mixed Features](#)

- Aras Selvi · Mohammad Reza Belbasi · Martin Haugh · Wolfram Wiesemann
- abstract@[open-review](#): Recent work has leveraged the popular distributionally robust optimization paradigm to combat overfitting in classical logistic regression. While the resulting classification scheme displays a promising performance in numerical experiments, it is inherently limited to numerical features. In this paper, we show that distributionally robust logistic regression with mixed (i.e., numerical and categorical) features, despite amounting to an optimization problem of exponential size, admits a polynomial-time solution scheme. We subsequently develop a practically efficient cutting plane approach that solves the problem as a sequence of polynomial-time solvable exponential conic programs. Our method retains many of the desirable theoretical features of previous works, but---in contrast to the literature---it does not admit an equivalent representation as a regularized logistic regression, that is, it represents a genuinely novel variant of the logistic regression problem. We show that our method outperforms both the unregularized and the regularized logistic regression on categorical as well as mixed-feature benchmark instances.

## [SecureFedYJ: a safe feature Gaussianization protocol for Federated Learning](#)

- Tanguy Marchand · Boris Muzellec · Constance Béguier · Jean Ogier du Terrail · Mathieu Andreux
- abstract@[open-review](#): The Yeo-Johnson (YJ) transformation is a standard parametrized per-feature unidimensional transformation often used to Gaussianize features in machine learning. In this paper, we investigate the problem of applying the YJ transformation in a cross-silo Federated Learning setting under privacy constraints. For the first time, we prove that the YJ negative log-likelihood is in fact convex, which allows us to optimize it with exponential search. We numerically show that the resulting algorithm is more stable than the state-of-the-art approach based on the Brent minimization method. Building on this simple algorithm and Secure Multiparty Computation routines, we propose SECUREFEDYJ, a federated algorithm that performs a pooled-equivalent YJ transformation without leaking more information than the final fitted parameters do. Quantitative experiments on real data demonstrate that, in addition to being secure, our approach reliably normalizes features across silos as well as if data were pooled, making it a viable approach for safe federated feature Gaussianization.

## [Smoothed Online Convex Optimization Based on Discounted-Normal-Predictor](#)

- Lijun Zhang · Wei Jiang · Jinfeng Yi · Tianbao Yang
- abstract@[open-review](#): In this paper, we investigate an online prediction strategy named as Discounted-Normal-Predictor (Kapralov and Panigrahy, 2010) for smoothed online convex optimization (SOCO), in which the learner needs to minimize not only the hitting cost but also the switching cost. In the setting of learning with expert advice, Daniely and Mansour (2019) demonstrate that Discounted-Normal-Predictor can be utilized to yield nearly optimal regret bounds over any interval, even in the presence of switching costs. Inspired by their results, we develop a simple algorithm for SOCO: Combining online gradient descent (OGD) with different step sizes sequentially by Discounted-Normal-Predictor. Despite its simplicity, we prove that it is able to minimize the adaptive regret with switching cost, i.e., attaining nearly optimal regret with switching cost on every interval. By exploiting the theoretical guarantee of OGD for dynamic regret, we further show that the proposed algorithm can minimize the dynamic regret with switching cost in every interval.

## [Online Decision Mediation from Scratch](#)

- Daniel Jarrett · Alihan Hacıyıldız · Mihaela van der Schaar
- abstract@[open-review](#): Consider learning a decision support assistant to serve as an intermediary between (oracle) expert behavior and (imperfect) human behavior: At each time, the algorithm observes an action chosen by a fallible agent, and decides whether to accept that agent's decision, intervene with an

alternative, or request the expert's opinion. For instance, in clinical diagnosis, fully-autonomous machine behavior is often beyond ethical affordances, thus real-world decision support is often limited to monitoring and forecasting. Instead, such an intermediary would strike a prudent balance between the former (purely prescriptive) and latter (purely descriptive) approaches, while providing an efficient interface between human mistakes and expert feedback. In this work, we first formalize the sequential problem of online decision mediation---that is, of simultaneously learning and evaluating mediator policies from scratch with abstentious feedback: In each round, deferring to the oracle obviates the risk of error, but incurs an upfront penalty, and reveals the otherwise hidden expert action as a new training data point. Second, we motivate and propose a solution that seeks to trade off (immediate) loss terms against (future) improvements in generalization error; in doing so, we identify why conventional bandit algorithms may fail. Finally, through experiments and sensitivities on a variety of datasets, we illustrate consistent gains over applicable benchmarks on performance measures with respect to the mediator policy, the learned model, and the decision-making system as a whole.

## [Transferring Pre-trained Multimodal Representations with Cross-modal Similarity Matching](#)

- Byoungjip Kim · Sungik Choi · Dasol Hwang · Moontae Lee · Honglak Lee
- abstract@[open-review](#): Despite surprising performance on zero-shot transfer, pre-training a large-scale multimodal model is often prohibitive as it requires a huge amount of data and computing resources. In this paper, we propose a method (BeamCLIP) that can effectively transfer the representations of a large pre-trained multimodal model (CLIP-ViT) into a small target model (e.g., ResNet-18). For unsupervised transfer, we introduce cross-modal similarity matching (CSM) that enables a student model to learn the representations of a teacher model by matching the relative similarity distribution across text prompt embeddings. To better encode the text prompts, we design context-based prompt augmentation (CPA) that can alleviate the lexical ambiguity of input text prompts. Our experiments show that unsupervised representation transfer of a pre-trained vision-language model enables a small ResNet-18 to achieve a better ImageNet-1K top-1 linear probe accuracy (66.2%) than vision-only self-supervised learning (SSL) methods (e.g., SimCLR: 51.8%, SwAV: 63.7%), while closing the gap with supervised learning (69.8%).

## [Dict-TTS: Learning to Pronounce with Prior Dictionary Knowledge for Text-to-Speech](#)

- Ziyue Jiang · Zhe Su · Zhou Zhao · Qian Yang · Yi Ren · Jinglin Liu · æŒ–æ³¼%o å ¶
- abstract@[open-review](#): Polyphone disambiguation aims to capture accurate pronunciation knowledge from natural text sequences for reliable Text-to-speech (TTS) systems. However, previous approaches require substantial annotated training data and additional efforts from language experts, making it difficult to extend high-quality neural TTS systems to out-of-domain daily conversations and countless languages worldwide. This paper tackles the polyphone disambiguation problem from a concise and novel perspective: we propose Dict-TTS, a semantic-aware generative text-to-speech model with an online website dictionary (the existing prior information in the natural language). Specifically, we design a semantics-to-pronunciation attention (S2PA) module to match the semantic patterns between the input text sequence and the prior semantics in the dictionary and obtain the corresponding pronunciations; The S2PA module can be easily trained with the end-to-end TTS model without any annotated phoneme labels. Experimental results in three languages show that our model outperforms several strong baseline models in terms of pronunciation accuracy and improves the prosody modeling of TTS systems. Further extensive analyses with different linguistic encoders demonstrate that each design in Dict-TTS is effective. Audio samples are available at <https://dicttts.github.io/DictTTS-Demo/>.

## [Rethinking Resolution in the Context of Efficient Video Recognition](#)

- Chuofan Ma · Qiushan Guo · Yi Jiang · Zehuan Yuan · Ping Luo · Xiaojuan Qi
- abstract@[open-review](#): In this paper, we empirically study how to make the most of low-resolution frames for efficient video recognition. Existing methods mainly focus on developing compact networks or alleviating temporal redundancy of video inputs to increase efficiency, whereas compressing frame resolution has rarely been considered a promising solution. A major concern with using low-resolution frames is its poor recognition accuracy. We thus start by analyzing the underlying causes of performance degradation on low-resolution frames. Our key finding is that, degradation in performance is not necessarily a result of degradation in input quality, but rather mismatch between network architecture and input scale. Motivated by the success of knowledge distillation (KD), we propose to bridge the gap between network and input size via cross-resolution KD (ResKD). Our work shows that ResKD is a simple but effective method to boost recognition accuracy on low-resolution frames. Without bells and whistles, ResKD considerably surpasses all competitive methods in terms of efficiency and accuracy on four large-scale benchmark datasets, i.e., ActivityNet, FCVID, Mini-Kinetics, Something-Something V2. In addition, we extensively demonstrate its effectiveness over state-of-the-art architectures, i.e., 3D-CNNs and Video Transformers, and scalability towards super low-resolution frames. The results suggest ResKD can serve as a general inference acceleration method for state-of-the-art video recognition.

## [Learning Energy Networks with Generalized Fenchel-Young Losses](#)

- Mathieu Blondel · Felipe Llinares-Lopez · Robert Dadashi · Leonard Hussonot · Matthieu Geist
- abstract@[open-review](#): Energy-based models, a.k.a. energy networks, perform inference by optimizing an energy function, typically parametrized by a neural network. This allows one to capture potentially complex relationships between inputs and outputs. To learn the parameters of the energy function, the solution to that optimization problem is typically fed into a loss function. The key challenge for training energy networks lies in computing loss gradients, as this typically requires argmin/argmax differentiation. In this paper, building upon a generalized notion of conjugate function, which replaces the usual bilinear pairing with a general energy function, we propose generalized Fenchel-Young losses, a natural loss construction for learning energy networks. Our losses enjoy many desirable properties and their gradients can be computed efficiently without argmin/argmax differentiation. We also prove the calibration of their excess risk in the case of linear-concave energies. We demonstrate our losses on multilabel classification and imitation learning tasks.

## [A Communication-Efficient Distributed Gradient Clipping Algorithm for Training Deep Neural Networks](#)

- Mingrui Liu · Zhenxun Zhuang · Yunwen Lei · Chunyang Liao
- abstract@[open-review](#): In distributed training of deep neural networks, people usually run Stochastic Gradient Descent (SGD) or its variants on each machine and communicate with other machines periodically. However, SGD might converge slowly in training some deep neural networks (e.g., RNN, LSTM) because of the exploding gradient issue. Gradient clipping is usually employed to address this issue in the single machine setting, but exploring this technique in the distributed setting is still in its infancy: it remains mysterious whether the gradient clipping scheme can take advantage of multiple machines to enjoy parallel speedup. The main technical difficulty lies in dealing with nonconvex loss function, non-Lipschitz continuous gradient, and skipping communication rounds simultaneously. In this paper, we explore a relaxed-smoothness assumption of the loss landscape which LSTM was shown to satisfy in previous works, and design a communication-efficient gradient clipping algorithm. This algorithm can be run on multiple machines, where each machine employs a gradient clipping scheme and communicate with other machines after multiple steps of gradient-based updates. Our algorithm is proved to have  $\mathcal{O}(\left(\frac{1}{N\epsilon^4}\right)^{\frac{1}{2}})$  iteration complexity and  $\mathcal{O}(\frac{1}{\epsilon^3})$  communication complexity for finding an  $\epsilon$ -stationary point in the homogeneous data setting, where  $N$  is the number of machines. This indicates that our algorithm enjoys linear speedup and reduced communication rounds. Our proof relies on novel analysis techniques of estimating truncated random variables, which we believe are of independent interest. Our experiments on several benchmark datasets and various scenarios demonstrate that our algorithm indeed exhibits fast convergence speed in practice and thus validates our theory.

## [A Deep Reinforcement Learning Framework for Column Generation](#)

- Cheng Chi · Amine Aboussalah · Elias Khalil · Juyoung Wang · Zoha Sherkat-Masoumi
- abstract@[open-review](#): Column Generation (CG) is an iterative algorithm for solving linear programs (LPs) with an extremely large number of variables (columns). CG is the workhorse for tackling large-scale integer linear programs, which rely on CG to solve LP relaxations within a branch and bound algorithm. Two canonical applications are the Cutting Stock Problem (CSP) and Vehicle Routing Problem with Time Windows (VRPTW). In VRPTW, for example, each binary variable represents the decision to include or exclude a route, of which there are exponentially many; CG incrementally grows the subset of columns being used, ultimately converging to an optimal solution. We propose RLCG, the first Reinforcement Learning (RL) approach for CG. Unlike typical column selection rules which myopically select a column based on local information at each iteration, we treat CG as a sequential decision-making problem, as the column selected in an iteration affects subsequent iterations of the algorithm. This perspective lends itself to a Deep Reinforcement Learning approach that uses Graph Neural Networks (GNNs) to represent the variable-constraint structure in the LP of interest. We perform an extensive set of experiments using the publicly available BPPLIB benchmark for CSP and Solomon benchmark for VRPTW. RLCG converges faster and reduces the number of CG iterations by 22.4% for CSP and 40.9% for VRPTW on average compared to a commonly used greedy policy.

## [Searching for Better Spatio-temporal Alignment in Few-Shot Action Recognition](#)

- Yichao Cao · Xiu Su · Qingfei Tang · Shan You · Xiaobo Lu · Chang Xu
- abstract@[open-review](#): Spatio-Temporal feature matching and alignment are essential for few-shot action recognition as they determine the coherence and effectiveness of the temporal patterns. Nevertheless, this process could be not reliable, especially when dealing with complex video scenarios. In this paper, we propose to improve the performance of matching and alignment from the end-to-end design of models. Our solution comes at two-folds. First, we encourage to enhance the extracted Spatio-Temporal representations from few-shot videos in the perspective of architectures. With this aim, we propose a specialized transformer search method for videos, thus the spatial and temporal attention can be well-organized and optimized for stronger feature representations. Second, we also design an efficient non-parametric spatio-temporal prototype alignment strategy to better handle the high variability of motion. In particular, a query-specific class prototype will be generated for each query sample and category, which can better match query sequences against all support sequences. By doing so, our method SST enjoys significant superiority over the benchmark UCF101 and HMDB51 datasets. For example, with no pretraining, our method achieves 17.1% Top-1 accuracy improvement than the baseline TRX on UCF101 5-way 1-shot setting but with only 3x fewer FLOPs.

## [Self-Aware Personalized Federated Learning](#)

- Huili Chen · Jie Ding · Eric W Tramel · Shuang Wu · Anit Kumar Sahu · Salman Avestimehr · Tao Zhang
- abstract@[open-review](#): In the context of personalized federated learning (FL), the critical challenge is to balance local model improvement and global model tuning when the personal and global objectives may not be exactly aligned. Inspired by Bayesian hierarchical models, we develop a self-aware personalized FL method where each client can automatically balance the training of its local personal model and the global model that implicitly contributes to other clients' training. Such a balance is derived from the inter-client and intra-client uncertainty quantification. A larger inter-client variation implies more personalization is needed. Correspondingly, our method uses uncertainty-driven local training steps an aggregation rule instead of conventional local fine-tuning and sample size-based aggregation. With experimental studies on synthetic data, Amazon Alexa audio data, and public datasets such as MNIST, FEMNIST, CIFAR10, and Sent140, we show that our proposed method can achieve significantly improved personalization performance compared with the existing counterparts.

## [Finite-Sample Maximum Likelihood Estimation of Location](#)

- Shivam Gupta · Jasper Lee · Eric Price · Paul Valiant
- abstract@[open-review](#): We consider 1-dimensional location estimation, where we estimate a parameter  $\lambda$  from  $n$  samples  $\lambda + \eta_i$ , with each  $\eta_i$  drawn i.i.d. from a known distribution  $f$ . For fixed  $f$  the maximum-likelihood estimate (MLE) is well-known to be optimal in the limit as  $n \rightarrow \infty$ : it is asymptotically normal with variance matching the Cramer-Rao lower bound of  $\frac{1}{n\mathcal{I}}$ , where  $\mathcal{I}$  is the Fisher information of  $f$ . However, this bound does not hold for finite  $n$ , or when  $f$  varies with  $n$ . We show for arbitrary  $f$  and  $n$  that one can recover a similar theory based on the Fisher information of a smoothed version of  $f$ , where the smoothing radius decays with  $n$ .

## [Thompson Sampling Efficiently Learns to Control Diffusion Processes](#)

- Mohamad Kazem Shirani Faradonbeh · Mohamad Sadegh Shirani Faradonbeh · Mohsen Bayati
- abstract@[open-review](#): Diffusion processes that evolve according to linear stochastic differential equations are an important family of continuous-time dynamic decision-making models. Optimal policies are well-studied for them, under full certainty about the drift matrices. However, little is known about data-driven control of diffusion processes with uncertain drift matrices as conventional discrete-time analysis techniques are not applicable. In addition, while the task can be viewed as a reinforcement learning problem involving exploration and exploitation trade-off, ensuring system stability is a fundamental component of designing optimal policies. We establish that the popular Thompson sampling algorithm learns optimal actions fast, incurring only a square-root of time regret, and also stabilizes the system in a short time period. To the best of our knowledge, this is the first such result for Thompson sampling in a diffusion process control problem. We validate our theoretical results through empirical simulations with real parameter matrices from two settings of airplane and blood glucose control. Moreover, we observe that Thompson sampling significantly improves (worst-case) regret, compared to the state-of-the-art algorithms, suggesting Thompson sampling explores in a more guarded fashion. Our theoretical analysis involves characterization of a certain optimality manifold that ties the local geometry of the drift parameters to the optimal control of the diffusion process. We expect this technique to be of broader interest.

## [Stability and Generalization for Markov Chain Stochastic Gradient Methods](#)

- Puyu Wang · Yunwen Lei · Yiming Ying · Ding-Xuan Zhou
- abstract@[open-review](#): Recently there is a large amount of work devoted to the study of Markov chain stochastic gradient methods (MC-SGMs) which mainly focus on their convergence analysis for solving minimization problems. In this paper, we provide a comprehensive generalization analysis of MC-SGMs for both minimization and minimax problems through the lens of algorithmic stability in the framework of statistical learning theory. For empirical risk minimization (ERM) problems, we establish the optimal excess population risk bounds for both smooth and non-smooth cases by introducing on-average argument stability. For minimax problems, we develop a quantitative connection between on-average argument stability and generalization error which extends the existing results for uniform stability (Lei et al., 2021). We further develop the first nearly optimal convergence rates for convex-concave problems both in expectation and with high probability, which, combined with our stability results, show that the optimal generalization bounds can be attained for both smooth and non-smooth cases. To the best of our knowledge, this is the first generalization analysis of SGMs when the gradients are sampled from a Markov process.

## [Transformers meet Stochastic Blockmodels: Attention with Data-Adaptive Sparsity and Cost](#)

- Sungjun Cho · Seonwoo Min · Jinwoo Kim · Moontae Lee · Honglak Lee · Seunghoon Hong
- abstract@[open-review](#): To overcome the quadratic cost of self-attention, recent works have proposed various sparse attention modules, most of which fall under one of two groups: 1) sparse attention under a hand-crafted patterns and 2) full attention followed by a sparse variant of softmax such as  $\alpha$ -entmax. Unfortunately, the first group lacks adaptability to data while the second still requires quadratic cost in training. In this work, we propose SBM-Transformer, a model that resolves both problems by endowing each attention head with a mixed-membership Stochastic Block Model (SBM). Then, each

attention head data-adaptively samples a bipartite graph, the adjacency of which is used as an attention mask for each input. During backpropagation, a straight-through estimator is used to flow gradients beyond the discrete sampling step and adjust the probabilities of sampled edges based on the predictive loss. The forward and backward cost are thus linear to the number of edges, which each attention head can also choose flexibly based on the input. By assessing the distribution of graphs, we theoretically show that SBM-Transformer is a universal approximator for arbitrary sequence-to-sequence functions in expectation. Empirical evaluations under the Long Range Arena benchmark demonstrate that our model outperforms previous efficient variants as well as the original Transformer with full attention.

## [FreGAN: Exploiting Frequency Components for Training GANs under Limited Data](#)

- mengping yang · Zhe Wang · Ziqiu Chi · Yanbing Zhang
- abstract@[open-review](#): Training GANs under limited data often leads to discriminator overfitting and memorization issues, causing divergent training. Existing approaches mitigate the overfitting by employing data augmentations, model regularization, or attention mechanisms. However, they ignore the frequency bias of GANs and take poor consideration towards frequency information, especially high-frequency signals that contain rich details. To fully utilize the frequency information of limited data, this paper proposes FreGAN, which raises the model's frequency awareness and draws more attention to synthesising high-frequency signals, facilitating high-quality generation. In addition to exploiting both real and generated images' frequency information, we also involve the frequency signals of real images as a self-supervised constraint, which alleviates the GAN disequilibrium and encourages the generator to synthesis adequate rather than arbitrary frequency signals. Extensive results demonstrate the superiority and effectiveness of our FreGAN in ameliorating generation quality in the low-data regime (especially when training data is less than 100). Besides, FreGAN can be seamlessly applied to existing regularization and attention mechanism models to further boost the performance.

## [Expansion and Shrinkage of Localization for Weakly-Supervised Semantic Segmentation](#)

- JINLONG LI · Zequn Jie · Xu Wang · Xiaolin Wei · Lin Ma
- abstract@[open-review](#): Generating precise class-aware pseudo ground-truths,  $\text{CAMs}$ , is essential for weakly-supervised semantic segmentation. The original CAM method usually produces incomplete and inaccurate localization maps. To tackle with this issue, this paper proposes an Expansion and Shrinkage scheme based on the offset learning in the deformable convolution, to sequentially improve the recall and precision of the located object in the two respective stages. In the Expansion stage, an offset learning branch in a deformable convolution layer, referred as expansion sampler" seeks for sampling increasingly less discriminative object regions<sup>1/4</sup>driven by an inverse supervision signal that maximizes image-level classification loss. The located more complete object in the Expansion stage is then gradually narrowed down to the final object region during the Shrinkage stage. In the Shrinkage stage, the offset learning branch of another deformable convolution layer, referred as shrinkage sampler", is introduced to exclude the false positive background regions attended in the Expansion stage to improve the precision of the localization maps. We conduct various experiments on PASCAL VOC 2012 and MS COCO 2014 to well demonstrate the superiority of our method over other state-of-the-art methods for weakly-supervised semantic segmentation.

## [Private Multiparty Perception for Navigation](#)

- Hui Lu · Mia Chiquier · Carl Vondrick
- abstract@[open-review](#): We introduce a framework for navigating through cluttered environments by connecting multiple cameras together while simultaneously preserving privacy. Occlusions and obstacles in large environments are often challenging situations for navigation agents because the environment is not fully observable from a single camera view. Given multiple camera views of an environment, our approach learns to produce a multiview scene representation that can only be used for navigation, provably preventing one party from inferring anything beyond the output task. On a new navigation dataset that we will publicly release, experiments show that private multiparty representations allow navigation through complex scenes and around obstacles while jointly preserving privacy. Our approach scales to an arbitrary number of camera viewpoints. We believe developing visual representations that preserve privacy is increasingly important for many applications such as navigation.

## [Global Convergence of Federated Learning for Mixed Regression](#)

- Lili Su · Jiaming Xu · Pengkun Yang
- abstract@[open-review](#): This paper studies the problem of model training under Federated Learning when clients exhibit cluster structure. We contextualize this problem in mixed regression, where each client has limited local data generated from one of  $k$  unknown regression models. We design an algorithm that achieves global convergence from any initialization, and works even when local data volume is highly unbalanced -- there could exist clients that contain  $O(1)$  data points only. Our algorithm first runs moment descent on a few anchor clients (each with  $\tilde{\Omega}(k)$  data points) to obtain coarse model estimates. Then each client alternately estimates its cluster labels and refines the model estimates based on FedAvg or FedProx. A key innovation in our analysis is a uniform estimate on the clustering errors, which we prove by bounding the VC dimension of general polynomial concept classes based on the theory of algebraic geometry.

## [A Unified Framework for Alternating Offline Model Training and Policy Learning](#)

- Shentao Yang · Yihao Feng · Shujian Zhang · Mingyuan Zhou
- abstract@[open-review](#): In offline model-based reinforcement learning (offline MBRL), we learn a dynamic model from historically collected data, and then utilize the learned model and fixed dataset for policy learning, without further interacting with the environment. Offline MBRL algorithms can improve the efficiency and stability of policy learning over the model-free based algorithms. However, in most of the existing offline MBRL algorithms, the learning objectives for the dynamic models and the policies are isolated from each other. Such an objective mismatch issue may lead to inferior performance of the learned agents. In this paper, we address the issue by developing an iterative offline MBRL framework, where we maximize a lower bound of the true expected return, by alternating between dynamic model training and policy learning. With the proposed unified model-policy learning framework, we achieve competitive performance on a wide range of continuous control reinforcement learning datasets.

## [Iron: Private Inference on Transformers](#)

- Meng Hao · Hongwei Li · Hanxiao Chen · Pengzhi Xing · Guowen Xu · Tianwei Zhang
- abstract@[open-review](#): We initiate the study of private inference on Transformer-based models in the client-server setting, where clients have private inputs and servers hold proprietary models. Our main contribution is to provide several new secure protocols for matrix multiplication and complex non-linear functions like Softmax, GELU activations, and LayerNorm, which are critical components of Transformers. Specifically, we first propose a customized homomorphic encryption-based protocol for matrix multiplication that crucially relies on a novel compact packing technique. This design achieves  $\sqrt{m} \times m$  less communication ( $m$  is the number of rows of the output matrix) and at least  $3 \times m^2$  performance improvement over the state-of-the-art. Second, we design efficient protocols for three non-linear functions via integrating advanced underlying protocols and specialized optimizations. Compared to the most efficient protocols, our recipes reduce about half of the communication and computation overhead. Furthermore, all protocols are numerically precise, which preserve the model accuracy of plaintext. These techniques together allow us to implement  $\text{Iron}$ , an efficient Transformer-based private inference framework. Experiments conducted on several real-world datasets and models demonstrate that  $\text{Iron}$  achieves  $3 \sim 14$  times less communication and  $3 \sim 11$  times less runtime compared to the prior art.

## [Towards Practical Few-shot Query Sets: Transductive Minimum Description Length Inference](#)

- SÃ©golÃ¨ne Martin Â· Malik Boudiaf Â· Emilie Chouzenoux Â· Jean-Christophe Pesquet Â· Ismail Ayed
- abstract@[open-review](#): Standard few-shot benchmarks are often built upon simplifying assumptions on the query sets, which may not always hold in practice. In particular, for each task at testing time, the classes effectively present in the unlabeled query set are known a priori, and correspond exactly to the set of classes represented in the labeled support set. We relax these assumptions and extend current benchmarks, so that the query-set classes of a given task are unknown, but just belong to a much larger set of possible classes. Our setting could be viewed as an instance of the challenging yet practical problem of extremely imbalanced K-way classification, K being much larger than the values typically used in standard benchmarks, and with potentially irrelevant supervision from the support set. Expectedly, our setting incurs drops in the performances of state-of-the-art methods. Motivated by these observations, we introduce a primal dual minimum description length (PADDLE) formulation, which balances data-fitting accuracy and model complexity for a given few-shot task, under supervision constraints from the support set. Our constrained MDL-like objective promotes competition among a large set of possible classes, preserving only effective classes that befit better the data of a few-shot task. It is hyper-parameter free, and could be applied on top of any base-class training. Furthermore, we derive a fast block coordinate descent algorithm for optimizing our objective, with convergence guarantee, and a linear computational complexity at each iteration. Comprehensive experiments over the standard few-shot datasets and the more realistic and challenging i-Nat dataset show highly competitive performances of our method, more so when the numbers of possible classes in the tasks increase.

## [Neural Differential Equations for Learning to Program Neural Nets Through Continuous Learning Rules](#)

- Kazuki Irie Â· Francesco Faccio Â· JÃ¼rgen Schmidhuber
- abstract@[open-review](#): Neural ordinary differential equations (ODEs) have attracted much attention as continuous-time counterparts of deep residual neural networks (NNs), and numerous extensions for recurrent NNs have been proposed. Since the 1980s, ODEs have also been used to derive theoretical results for NN learning rules, e.g., the famous connection between Oja's rule and principal component analysis. Such rules are typically expressed as additive iterative update processes which have straightforward ODE counterparts. Here we introduce a novel combination of learning rules and Neural ODEs to build continuous-time sequence processing nets that learn to manipulate short-term memory in rapidly changing synaptic connections of other nets. This yields continuous-time counterparts of Fast Weight Programmers and linear Transformers. Our novel models outperform the best existing Neural Controlled Differential Equation based models on various time series classification tasks, while also addressing their scalability limitations.

## [What I Cannot Predict, I Do Not Understand: A Human-Centered Evaluation Framework for Explainability Methods](#)

- Julien Colin Â· Thomas FEL Â· Remi Cadene Â· Thomas Serre
- abstract@[open-review](#): A multitude of explainability methods and theoretical evaluation scores have been proposed. However, it is not yet known: (1) how useful these methods are in real-world scenarios and (2) how well theoretical measures predict the usefulness of these methods for practical use by a human. To fill this gap, we conducted human psychophysics experiments at scale to evaluate the ability of human participants (n=1,150) to leverage representative attribution methods to predicting the decision of different image classifiers. We carried out this analysis on 3 datasets, each dedicated to one of the end-goal of explainability: bias detection, identification of new strategies and the understanding of failure cases. Our results demonstrate that the degree to which individual attribution methods helped human participants better understand a model varied widely across categorization tasks and datasets. We test several hypotheses to understand the reasons for these failures by investigating: (1) the relationship between explanation fidelity and usefulness, (2) the effects of explanation complexity, and (3) the prediction of usefulness using a human visual perceptual similarity proxy. Overall, our results highlight fundamental challenges for the field -- suggesting a critical need to refocus the development of explainability tools that go beyond attribution methods. We will make the code of our framework and results available to ease the systematic evaluation of novel explainability methods and to support the development of theoretical measures more aligned with human.

## [QueryPose: Sparse Multi-Person Pose Regression via Spatial-Aware Part-Level Query](#)

- Yabo Xiao Â· Xiaojuan Wang Â· Kai Su Â· Dongdong Yu Â· Lei Jin Â· Mingshu He Â· Zehuan Yuan
- abstract@[open-review](#): We propose a sparse end-to-end multi-person pose regression framework, termed QueryPose, which can directly predict multi-person keypoint sequences from the input image. The existing end-to-end methods rely on dense representations to preserve the spatial detail and structure for precise keypoint localization. However, the dense paradigm introduces complex and redundant post-processes during inference. In our framework, each human instance is encoded by several learnable spatial-aware part-level queries associated with an instance-level query. First, we propose the Spatial Part Embedding Generation Module (SPEGM) that considers the local spatial attention mechanism to generate several spatial-sensitive part embeddings, which contain spatial details and structural information for enhancing the part-level queries. Second, we introduce the Selective Iteration Module (SIM) to adaptively update the sparse part-level queries via the generated spatial-sensitive part embeddings stage-by-stage. Based on the two proposed modules, the part-level queries are able to fully encode the spatial details and structural information for precise keypoint regression. With the bipartite matching, QueryPose avoids the hand-designed post-processes. Without bells and whistles, QueryPose surpasses all existing dense end-to-end methods with 73.6 AP on MS COCO mini-val set and 72.7 AP on CrowdPose test set. Code will be released.

## [KERPLE: Kernelized Relative Positional Embedding for Length Extrapolation](#)

- Ta-Chung Chi Â· Ting-Han Fan Â· Peter J Ramadge Â· Alexander Rudnick
- abstract@[open-review](#): Relative positional embeddings (RPE) have received considerable attention since RPEs effectively model the relative distance among tokens and enable length extrapolation. We propose KERPLE, a framework that generalizes relative position embedding for extrapolation by kernelizing positional differences. We achieve this goal using conditionally positive definite (CPD) kernels, a class of functions known for generalizing distance metrics. To maintain the inner product interpretation of self-attention, we show that a CPD kernel can be transformed into a PD kernel by adding a constant offset. This offset is implicitly absorbed in the Softmax normalization during self-attention. The diversity of CPD kernels allows us to derive various RPEs that enable length extrapolation in a principled way. Experiments demonstrate that the logarithmic variant achieves excellent extrapolation performance on three large language modeling datasets.

## [Cross-Image Context for Single Image Inpainting](#)

- Tingliang Feng Â· Wei Feng Â· Weiqi Li Â· Di Lin
- abstract@[open-review](#): Visual context is of crucial importance for image inpainting. The contextual information captures the appearance and semantic correlation between the image regions, helping to propagate the information of the complete regions for reasoning the content of the corrupted regions. Many inpainting methods compute the visual context based on the regions within the single image. In this paper, we propose the Cross-Image Context Memory (CICM) for learning and using the cross-image context to recover the corrupted regions. CICM consists of multiple sets of the cross-image representations learned from the image regions with different visual patterns. The regional representations are learned across different images, thus providing richer context that benefit the inpainting task. The experimental results demonstrate the effectiveness and generalization of CICM, which achieves state-of-the-art performances on various datasets for single image inpainting.

## [Deep Learning Methods for Proximal Inference via Maximum Moment Restriction](#)

- Benjamin Kompa Â· David Bellamy Â· Tom Kolokotrones Â· James M Robins Â· Andrew Beam

- abstract@[open-review](#): The No Unmeasured Confounding Assumption is widely used to identify causal effects in observational studies. Recent work on proximal inference has provided alternative identification results that succeed even in the presence of unobserved confounders, provided that one has measured a sufficiently rich set of proxy variables, satisfying specific structural conditions. However, proximal inference requires solving an ill-posed integral equation. Previous approaches have used a variety of machine learning techniques to estimate a solution to this integral equation, commonly referred to as the bridge function. However, prior work has often been limited by relying on pre-specified kernel functions, which are not data adaptive and struggle to scale to large datasets. In this work, we introduce a flexible and scalable method based on a deep neural network to estimate causal effects in the presence of unmeasured confounding using proximal inference. Our method achieves state of the art performance on two well-established proximal inference benchmarks. Finally, we provide theoretical consistency guarantees for our method.

## [InsPro: Propagating Instance Query and Proposal for Online Video Instance Segmentation](#)

- Fei He · Naiyu Gao · Jian Jia · Haoyang Zhang · Yanhu Shan · Xin Zhao · Kaiqi Huang
- abstract@[open-review](#): Video instance segmentation (VIS) aims at segmenting and tracking objects in videos. Prior methods typically first generate frame-level or clip-level object instances and then associate them by either additional tracking heads or complex instance matching algorithms. This explicit instance association approach increases system complexity and fails to fully exploit temporal cues in videos. In this paper, we design a simple, fast and yet effective query-based framework for online VIS. Relying on an instance query and proposal propagation mechanism with several specially developed components, this framework can perform accurate instance association implicitly. Specifically, we generate frame-level object instances based on a set of instance query-proposal pairs propagated from previous frames. This instance query-proposal pair is learned to bind with one specific object across frames through conscientiously developed strategies. When using such a pair to predict an object instance on the current frame, not only the generated instance is automatically associated with its precursors on previous frames, but the model gets a good prior for predicting the same object. In this way, we naturally achieve implicit instance association in parallel with segmentation and elegantly take advantage of temporal clues in videos. To show the effectiveness of our method InsPro, we evaluate it on two popular VIS benchmarks, i.e., YouTube-VIS 2019 and YouTube-VIS 2021. Without bells-and-whistles, our InsPro with ResNet-50 backbone achieves 43.2 AP and 37.6 AP on these two benchmarks respectively, outperforming all other online VIS methods. Code will be made publicly available.

## [GET3D: A Generative Model of High Quality 3D Textured Shapes Learned from Images](#)

- Jun Gao · Tianchang Shen · Zian Wang · Wenzheng Chen · Kangxue Yin · Daiqing Li · Or Litany · Zan Gojcic · Sanja Fidler
- abstract@[open-review](#): As several industries are moving towards modeling massive 3D virtual worlds, the need for content creation tools that can scale in terms of the quantity, quality, and diversity of 3D content is becoming evident. In our work, we aim to train performant 3D generative models that synthesize textured meshes which can be directly consumed by 3D rendering engines, thus immediately usable in downstream applications. Prior works on 3D generative modeling either lack geometric details, are limited in the mesh topology they can produce, typically do not support textures, or utilize neural renderers in the synthesis process, which makes their use in common 3D software non-trivial. In this work, we introduce GET3D, a Generative model that directly generates Explicit Textured 3D meshes with complex topology, rich geometric details, and high fidelity textures. We bridge recent success in the differentiable surface modeling, differentiable rendering as well as 2D Generative Adversarial Networks to train our model from 2D image collections. GET3D is able to generate high-quality 3D textured meshes, ranging from cars, chairs, animals, motorbikes and human characters to buildings, achieving significant improvements over previous methods.

## [On the generalization of learning algorithms that do not converge](#)

- Nisha Chandramoorthy · Andreas Loukas · Khashayar Gatmiry · Stefanie Jegelka
- abstract@[open-review](#): Common generalization analyses of deep learning assume that the training converges to a fixed point. But, in practice, the weights of deep neural networks optimized with stochastic gradient descent often do not exhibit convergent behavior. To reduce this discrepancy between theory and practice, we analyze the generalization of algorithms when the weights only converge in distribution. Our main contribution is to propose a notion of "statistical algorithmic stability" (SAS) that extends classical algorithmic stability to non-convergent algorithms and to study its connection to generalization. This ergodic-theoretic approach to learning theory leads to new insights when compared to the traditional optimization and learning theory perspectives. We prove that the stability of the time-asymptotic behavior of a learning algorithm can be used to predict its generalization and empirically demonstrate how loss dynamics can provide clues to generalization performance. Our findings provide evidence that networks that "train stably generalize better" even when the training continues indefinitely and the weights do not converge.

## [LobsDICE: Offline Learning from Observation via Stationary Distribution Correction Estimation](#)

- Geon-Hyeong Kim · Jongmin Lee · Youngsoo Jang · Hongseok Yang · Kee-Eung Kim
- abstract@[open-review](#): We consider the problem of learning from observation (LfO), in which the agent aims to mimic the expert's behavior from the state-only demonstrations by experts. We additionally assume that the agent cannot interact with the environment but has access to the action-labeled transition data collected by some agents with unknown qualities. This offline setting for LfO is appealing in many real-world scenarios where the ground-truth expert actions are inaccessible and the arbitrary environment interactions are costly or risky. In this paper, we present LobsDICE, an offline LfO algorithm that learns to imitate the expert policy via optimization in the space of stationary distributions. Our algorithm solves a single convex minimization problem, which minimizes the divergence between the two state-transition distributions induced by the expert and the agent policy. Through an extensive set of offline LfO tasks, we show that LobsDICE outperforms strong baseline methods.

## [End-to-end Symbolic Regression with Transformers](#)

- Pierre-alexandre Kamienny · Stéphane d'Ascoli · Guillaume Lample · François Fleuret
- abstract@[open-review](#): Symbolic regression, the task of predicting the mathematical expression of a function from the observation of its values, is a difficult task which usually involves a two-step procedure: predicting the "skeleton" of the expression up to the choice of numerical constants, then fitting the constants by optimizing a non-convex loss function. The dominant approach is genetic programming, which evolves candidates by iterating this subroutine a large number of times. Neural networks have recently been tasked to predict the correct skeleton in a single try, but remain much less powerful. In this paper, we challenge this two-step procedure, and task a Transformer to directly predict the full mathematical expression, constants included. One can subsequently refine the predicted constants by feeding them to the non-convex optimizer as an informed initialization. We present ablations to show that this end-to-end approach yields better results, sometimes even without the refinement step. We evaluate our model on problems from the SRBench benchmark and show that our model approaches the performance of state-of-the-art genetic programming with several orders of magnitude faster inference.

## [Teach Less, Learn More: On the Undistillable Classes in Knowledge Distillation](#)

- Yichen Zhu · Ning Liu · Zhiyuan Xu · Xin Liu · Weibin Meng · Louis Wang · Zhicai Ou · Jian Tang
- abstract@[open-review](#): Knowledge distillation (KD) can effectively compress neural networks by training a smaller network (student) to simulate the behavior of a larger one (teacher). A counter-intuitive observation is that a more expansive teacher does not make a better student, but the reasons for this phenomenon remain unclear. In this paper, we demonstrate that this is directly attributed to the presence of \textit{undistillable classes}: when trained with distillation, the teacher's knowledge of some classes is incomprehensible to the student model. We observe that while KD improves the overall accuracy, it is at the cost of the model becoming inaccurate in these undistillable classes. After establishing their widespread existence in state-of-the-art distillation

methods, we illustrate their correlation with the capacity gap between teacher and student models. Finally, we present a simple Teach Less Learn More (TLLM) framework to identify and discard the undistillable classes during training. We validate the effectiveness of our approach on multiple datasets with varying network architectures. In all settings, our proposed method is able to exceed the performance of competitive state-of-the-art techniques.

## [Hypothesis Testing for Differentially Private Linear Regression](#)

- Daniel Alabi · Salil Vadhan
- abstract@[open-review](#): In this work, we design differentially private hypothesis tests for the following problems in the general linear model: testing a linear relationship and testing for the presence of mixtures. The majority of our hypothesis tests are based on differentially private versions of the  $\$F\$$ -statistic for the general linear model framework, which are uniformly most powerful unbiased in the non-private setting. We also present another test for testing mixtures, based on the differentially private nonparametric tests of Couch, Kazan, Shi, Bray, and Groce (CCS 2019), which is especially suited for the small dataset regime. We show that the differentially private  $\$F\$$ -statistic converges to the asymptotic distribution of its non-private counterpart. As a corollary, the statistical power of the differentially private  $\$F\$$ -statistic converges to the statistical power of the non-private  $\$F\$$ -statistic. Through a suite of Monte Carlo based experiments, we show that our tests achieve desired \textit{significance levels} and have a high \textit{power} that approaches the power of the non-private tests as we increase sample sizes or the privacy-loss parameter. We also show when our tests outperform existing methods in the literature.

## [Let Images Give You More: Point Cloud Cross-Modal Training for Shape Analysis](#)

- Xu Yan · Heshen Zhan · Chaoda Zheng · Jiantao Gao · Ruimao Zhang · Shuguang Cui · Zhen Li
- abstract@[open-review](#): Although recent point cloud analysis achieves impressive progress, the paradigm of representation learning from single modality gradually meets its bottleneck. In this work, we take a step towards more discriminative 3D point cloud representation using 2D images, which inherently contain richer appearance information, e.g., texture, color, and shade. Specifically, this paper introduces a simple but effective point cloud cross-modality training (PointCMT) strategy, which utilizes view-images, i.e., rendered or projected 2D images of the 3D object, to boost point cloud classification. In practice, to effectively acquire auxiliary knowledge from view-images, we develop a teacher-student framework and formulate the cross-modal learning as a knowledge distillation problem. Through novel feature and classifier enhancement criteria, PointCMT eliminates the distribution discrepancy between different modalities and avoid potential negative transfer effectively. Note that PointCMT efficiently improves the point-only representation without any architecture modification. Sufficient experiments verify significant gains on various datasets based on several backbones, i.e., equipped with PointCMT, PointNet++ and PointMLP achieve state-of-the-art performance on two benchmarks, i.e., 94.4% and 86.7% accuracy on ModelNet40 and ScanObjectNN, respectively.

## [Contextual Dynamic Pricing with Unknown Noise: Explore-then-UCB Strategy and Improved Regrets](#)

- Yiyun Luo · Will Wei Sun · Yufeng Liu
- abstract@[open-review](#): Dynamic pricing is a fast-moving research area in machine learning and operations management. A lot of work has been done for this problem with known noise. In this paper, we consider a contextual dynamic pricing problem under a linear customer valuation model with an unknown market noise distribution  $\$F\$$ . This problem is very challenging due to the difficulty in balancing three tangled tasks of revenue-maximization, estimating the linear valuation parameter  $\$\\theta_0\$$ , and learning the nonparametric  $\$F\$$ . To address this issue, we develop a novel \{it Explore-then-UCB\} (ExUCB) strategy that includes an exploration for  $\$\\theta_0\$$ -learning and a followed UCB procedure of joint revenue-maximization and  $\$F\$$ -learning. Under Lipschitz and 2nd-order smoothness assumptions on  $\$F\$$ , ExUCB is the first approach to achieve the  $\$\\tilde{O}(T^{2/3})\$$  regret rate. Under the Lipschitz assumption only, ExUCB matches the best existing regret of  $\$\\tilde{O}(T^{3/4})\$$  and is computationally more efficient. Furthermore, for regret lower bounds under the nonparametric  $\$F\$$ , not much work has been done beyond only assuming Lipschitz. To fill this gap, we provide the first  $\$\\tilde{\\Omega}(T^{3/5})\$$  lower bound under Lipschitz and 2nd-order smoothness assumptions.

## [ClimbQ: Class Imbalanced Quantization Enabling Robustness on Efficient Inferences](#)

- Ting-An Chen · Ming-syan Chen
- abstract@[open-review](#): Quantization compresses models to low bits for efficient inferences which has received increasing attentions. However, existing approaches focused on balanced datasets, while imbalanced data is pervasive in the real world. Therefore, in this study, we investigate the realistic problem, quantization on class-imbalanced data. We observe from the analytical results that quantizing imbalanced data inclines to obtain a large error due to the differences between separate class distributions, which leads to a significant accuracy loss. To address this issue, we propose a novel quantization framework, Class Imbalanced Quantization (ClimbQ) that focuses on diminishing the inter-class heterogeneity for quantization error reduction. ClimbQ first scales the variance of each class distribution and then projects data through the new distributions to the same space for quantization. To guarantee the homogeneity of class variances after the ClimbQ process, we examine the quantized features and derive that the homogeneity satisfies when data size for each class is restricted (bounded). Accordingly, we design a Homogeneous Variance Loss (HomoVar Loss) which reweights the data losses of each class based on the bounded data sizes to satisfy the homogeneity of class variances. Extensive experiments on class-imbalanced and benchmark balanced datasets reveal that ClimbQ outperforms the state-of-the-art quantization techniques, especially on highly imbalanced data.

## [Sparse2Dense: Learning to Densify 3D Features to Boost 3D Object Detection](#)

- Tianyu Wang · Xiaowei Hu · Zhenghe LIU · Chi-Wing Fu
- abstract@[open-review](#): LiDAR-produced point clouds are the major source for most state-of-the-art 3D object detectors. Yet, small, distant, and incomplete objects with sparse or few points are often hard to detect. We present Sparse2Dense, a new framework to efficiently boost 3D detection performance by learning to densify point clouds in latent space. Specifically, we first train a dense point 3D detector (DDet) with a dense point cloud as input and design a sparse point 3D detector (SDet) with a regular point cloud as input. Importantly, we formulate the lightweight plug-in S2D module and the point cloud reconstruction module in SDet to densify 3D features and train SDet to produce 3D features, following the dense 3D features in DDet. So, in inference, SDet can simulate dense 3D features from regular (sparse) point cloud inputs without requiring dense inputs. We evaluate our method on the large-scale Waymo Open Dataset and the Waymo Domain Adaptation Dataset, showing its high performance and efficiency over the state of the arts.

## [Blackbox Attacks via Surrogate Ensemble Search](#)

- Zikui Cai · Srikanth Krishnamurthy · Chengyu Song · Amit Roy-Chowdhury · Salman Asif
- abstract@[open-review](#): Blackbox adversarial attacks can be categorized into transfer- and query-based attacks. Transfer methods do not require any feedback from the victim model, but provide lower success rates compared to query-based methods. Query attacks often require a large number of queries for success. To achieve the best of both approaches, recent efforts have tried to combine them, but still require hundreds of queries to achieve high success rates (especially for targeted attacks). In this paper, we propose a novel method for blackbox attacks via surrogate ensemble search (BASES) that can generate highly successful blackbox attacks using an extremely small number of queries. We first define a perturbation machine that generates a perturbed image by minimizing a weighted loss function over a fixed set of surrogate models. To generate an attack for a given victim model, we search over the weights in the loss function using queries generated by the perturbation machine. Since the dimension of the search space is small (same as the number of surrogate models), the search requires a small number of queries. We demonstrate that our proposed method achieves better success rate with at least \$30\times\$ fewer queries compared to state-of-the-art methods on different image classifiers trained with ImageNet (including VGG-19, DenseNet-121, and ResNext-50). In particular, our method requires as few as 3 queries per image (on average) to achieve more than a 90% success rate for targeted

attacks and 1--2 queries per image for over a 99% success rate for untargeted attacks. Our method is also effective on Google Cloud Vision API and achieved a 91% untargeted attack success rate with 2.9 queries per image. We also show that the perturbations generated by our proposed method are highly transferable and can be adopted for hard-label blackbox attacks.

## [A Statistical Online Inference Approach in Averaged Stochastic Approximation](#)

- Chuhan Xie · Zhihua Zhang
- abstract@[open-review](#): In this paper we propose a general framework to perform statistical online inference in a class of constant step size stochastic approximation (SA) problems, including the well-known stochastic gradient descent (SGD) and Q-learning. Regarding a constant step size SA procedure as a time-homogeneous Markov chain, we establish a functional central limit theorem (FCLT) for it under weaker conditions, and then construct confidence intervals for parameters via random scaling. To leverage the FCLT results in the Markov chain setting, an alternative condition that is more applicable for SA problems is established. We conduct experiments to perform inference with both random scaling and other traditional inference methods, and finds that the former has a more accurate and robust performance.

## [Predicting Label Distribution from Multi-label Ranking](#)

- Yunan Lu · Xiuyi Jia
- abstract@[open-review](#): Label distribution can provide richer information about label polysemy than logical labels in multi-label learning. There are currently two strategies including LDL (label distribution learning) and LE (label enhancement) to predict label distributions. LDL requires experts to annotate instances with label distributions and learn a predictive mapping on such a training set. LE requires experts to annotate instances with logical labels and generates label distributions from them. However, LDL requires costly annotation, and the performance of the LE is unstable. In this paper, we study the problem of predicting label distribution from multi-label ranking which is a compromise w.r.t. annotation cost but has good guarantees for performance. On the one hand, we theoretically investigate the relation between multi-label ranking and label distribution. We define the notion of EAE (expected approximation error) to quantify the quality of an annotation, give the bounds of EAE for multi-label ranking, and derive the optimal range of label distribution corresponding to a particular multi-label ranking. On the other hand, we propose a framework of label distribution predicting from multi-label ranking via conditional Dirichlet mixtures. This framework integrates the processes of recovering and learning label distributions end-to-end and allows us to easily encode our knowledge about current tasks by a scoring function. Finally, we implement extensive experiments to validate our proposal.

## [Improved Coresets for Euclidean \$k\$ -Means](#)

- Vincent Cohen-Addad · Kasper Green Larsen · David Saulpic · Chris Schwiegelshohn · Omar Ali Sheikh-Omar
- abstract@[open-review](#): Given a set of  $n$  points in  $d$  dimensions, the Euclidean  $k$ -means problem consists of finding  $k$  centers such that the sum of squared distances from every point to its closest center is minimized. The arguably most popular way of dealing with this problem in the big data setting is to first compress the data by computing a weighted subset known as a coreset and then run any algorithm on this subset. The guarantee of the coreset is that for any candidate solution, the ratio between coreset cost and the cost of the original instance is less than  $(1 + \epsilon)$  factor. The current state of the art coreset size for Euclidean  $k$ -means is  $\tilde{O}(k \cdot \epsilon^{-2} \cdot \min(k, \epsilon^{-2}))$ . This matches the lower bound of  $\Omega(k \cdot \epsilon^{-2})$  up to the  $\min(k, \epsilon^{-2})$  term. In this paper, we improve this bound to  $\min(\sqrt{k}, \epsilon^{-2})$ . In the regime where  $k \leq \epsilon^{-2}$ , this is a strict improvement over the state of the art. In particular, ours is the first provable bound that breaks through the  $k^2$  barrier while retaining an optimal dependency on  $\epsilon$ .

## [An Asymptotically Optimal Batched Algorithm for the Dueling Bandit Problem](#)

- Arpit Agarwal · Rohan Ghuge · viswanath nagarajan
- abstract@[open-review](#): We study the  $K$ -armed dueling bandit problem, a variation of the traditional multi-armed bandit problem in which feedback is obtained in the form of pairwise comparisons. Previous learning algorithms have focused on the fully adaptive setting, where the algorithm can make updates after every comparison. The "batched" dueling bandit problem is motivated by large-scale applications like web search ranking and recommendation systems, where performing sequential updates may be infeasible. In this work, we ask: is there a solution using only a few adaptive rounds that matches the asymptotic regret bounds of the best sequential algorithms for  $K$ -armed dueling bandits? We answer this in the affirmative under the Condorcet condition, a standard setting of the  $K$ -armed dueling bandit problem. We obtain asymptotic regret of  $O(K^2 \log^2(K)) + O(K \log(T))$  in  $O(\log(T))$  rounds, where  $T$  is the time horizon. Our regret bounds nearly match the best regret bounds known in the fully sequential setting under the Condorcet condition. Finally, in computational experiments over a variety of real-world datasets, we observe that our algorithm using  $O(\log(T))$  rounds achieves almost the same performance as fully sequential algorithms (that use  $T$  rounds).

## [Collaborative Learning of Distributions under Heterogeneity and Communication Constraints](#)

- Xinmeng Huang · Donghwan Lee · Edgar Dobriban · Hamed Hassani
- abstract@[open-review](#): In modern machine learning, users often have to collaborate to learn distributions that generate the data. Communication can be a significant bottleneck. Prior work has studied homogeneous users---i.e., whose data follow the same discrete distribution---and has provided optimal communication-efficient methods. However, these methods rely heavily on homogeneity, and are less applicable in the common case when users' discrete distributions are heterogeneous. Here we consider a natural and tractable model of heterogeneity, where users' discrete distributions only vary sparsely, on a small number of entries. We propose a novel two-stage method named SHIFT: First, the users collaborate by communicating with the server to learn a central distribution; relying on methods from robust statistics. Then, the learned central distribution is fine-tuned to estimate the individual distributions of users. We show that our method is minimax optimal in our model of heterogeneity and under communication constraints. Further, we provide experimental results using both synthetic data and  $n$ -gram frequency estimation in the text domain, which corroborate its efficiency.

## [Latent Hierarchical Causal Structure Discovery with Rank Constraints](#)

- Biwei Huang · Charles Jia Han Low · Feng Xie · Clark Glymour · Kun Zhang
- abstract@[open-review](#): Most causal discovery procedures assume that there are no latent confounders in the system, which is often violated in real-world problems. In this paper, we consider a challenging scenario for causal structure identification, where some variables are latent and they may form a hierarchical graph structure to generate the measured variables; the children of latent variables may still be latent and only leaf nodes are measured, and moreover, there can be multiple paths between every pair of variables (i.e., it is beyond tree structure). We propose an estimation procedure that can efficiently locate latent variables, determine their cardinalities, and identify the latent hierarchical structure, by leveraging rank deficiency constraints over the measured variables. We show that the proposed algorithm can find the correct Markov equivalence class of the whole graph asymptotically under proper restrictions on the graph structure and with linear causal relations.

## [Distributed Influence-Augmented Local Simulators for Parallel MARL in Large Networked Systems](#)

- Miguel Suau de Castro · Jinke He · Mustafa Mert Åzelikok · Matthijs Spaan · Frans Oliehoek
- abstract@[open-review](#): Due to its high sample complexity, simulation is, as of today, critical for the successful application of reinforcement learning. Many real-world problems, however, exhibit overly complex dynamics, which makes their full-scale simulation computationally slow. In this paper, we

show how to decompose large networked systems of many agents into multiple local components such that we can build separate simulators that run independently and in parallel. To monitor the influence that the different local components exert on one another, each of these simulators is equipped with a learned model that is periodically trained on real trajectories. Our empirical results reveal that distributing the simulation among different processes not only makes it possible to train large multi-agent systems in just a few hours but also helps mitigate the negative effects of simultaneous learning.

## [Uncovering the Structural Fairness in Graph Contrastive Learning](#)

- Ruijia Wang · Xiao Wang · Chuan Shi · Le Song
- abstract@[open-review](#): Graph contrastive learning (GCL) marries the power of graph convolutional network (GCN) and contrastive learning, which has emerged as a promising self-supervised approach for learning node representations. There is a basic demand for structural fairness in node representation learning, i.e., good performance on both low- and high-degree nodes. Recent studies illustrate that GCN often performs worse predictive accuracy for low-degree nodes, exhibiting the structural unfairness for prevalent long-tailed graphs. However, there is no literature exploring how GCL behaves with respect to node degree. In this work, we surprisingly find out that representations obtained by GCL methods are already fairer to degree bias than those learned by GCN. Then we theoretically prove this fairness stems from intra-community concentration and inter-community scatter properties of GCL, resulting in a much clear community structure to drive low-degree nodes away from the community boundary. Based on our theoretical analysis, we further devise a GRAph contrastive learning for DEgree bias (GRADE) based on a novel graph augmentation that applies different strategies to low- and high-degree nodes. Extensive experiments on various benchmarks and evaluation protocols validate the effectiveness of the proposed model.

## [SizeShiftReg: a Regularization Method for Improving Size-Generalization in Graph Neural Networks](#)

- Davide Buffelli · Pietro Lī · Fabio Vandin
- abstract@[open-review](#): In the past few years, graph neural networks (GNNs) have become the de facto model of choice for graph classification. While, from the theoretical viewpoint, most GNNs can operate on graphs of any size, it is empirically observed that their classification performance degrades when they are applied on graphs with sizes that differ from those in the training data. Previous works have tried to tackle this issue in graph classification by providing the model with inductive biases derived from assumptions on the generative process of the graphs, or by requiring access to graphs from the test domain. The first strategy is tied to the use of ad-hoc models and to the quality of the assumptions made on the generative process, leaving open the question of how to improve the performance of generic GNN models in general settings. On the other hand, the second strategy can be applied to any GNN, but requires access to information that is not always easy to obtain. In this work we consider the scenario in which we only have access to the training data, and we propose a regularization strategy that can be applied to any GNN to improve its generalization capabilities from smaller to larger graphs without requiring access to the test data. Our regularization is based on the idea of simulating a shift in the size of the training graphs using coarsening techniques, and enforcing the model to be robust to such a shift. Experimental results on standard datasets show that popular GNN models, trained on the 50% smallest graphs in the dataset and tested on the 10% largest graphs, obtain performance improvements of up to 30% when trained with our regularization strategy.

## [Enhance the Visual Representation via Discrete Adversarial Training](#)

- Xiaofeng Mao · YueFeng Chen · Gege Qi · Xiaodan Li · Ranjie Duan · Yao Zhu · shaokai ye · Rong Zhang · Hui Xue'
- abstract@[open-review](#): Adversarial Training (AT), which is commonly accepted as one of the most effective approaches defending against adversarial examples, can largely harm the standard performance, thus has limited usefulness on industrial-scale production and applications. Surprisingly, this phenomenon is totally opposite in Natural Language Processing (NLP) task, where AT can even benefit for generalization. We notice the merit of AT in NLP tasks could derive from the discrete and symbolic input space. For borrowing the advantage from NLP-style AT, we propose Discrete Adversarial Training (DAT). DAT leverages VQGAN to reform the image data to discrete text-like inputs, i.e. visual words. Then it minimizes the maximal risk on such discrete images with symbolic adversarial perturbations. We further give an explanation from the perspective of distribution to demonstrate the effectiveness of DAT. As a plug-and-play technique for enhancing the visual representation, DAT achieves significant improvement on multiple tasks including image classification, object detection and self-supervised learning. Especially, the model pre-trained with Masked Auto-Encoding (MAE) and fine-tuned by our DAT without extra data can get 31.40 mCE on ImageNet-C and 32.77% top-1 accuracy on Stylized-ImageNet, building the new state-of-the-art.

## [Boosting Out-of-distribution Detection with Typical Features](#)

- Yao Zhu · YueFeng Chen · Chuanlong Xie · Xiaodan Li · Rong Zhang · Hui Xue' · Xiang Tian · bolun zheng · Yaowu Chen
- abstract@[open-review](#): Out-of-distribution (OOD) detection is a critical task for ensuring the reliability and safety of deep neural networks in real-world scenarios. Different from most previous OOD detection methods that focus on designing OOD scores or introducing diverse outlier examples to retrain the model, we delve into the obstacle factors in OOD detection from the perspective of typicality and regard the feature's high-probability region of the deep model as the feature's typical set. We propose to rectify the feature into its typical set and calculate the OOD score with the typical features to achieve reliable uncertainty estimation. The feature rectification can be conducted as a plug-and-play module with various OOD scores. We evaluate the superiority of our method on both the commonly used benchmark (CIFAR) and the more challenging high-resolution benchmark with large label space (ImageNet). Notably, our approach outperforms state-of-the-art methods by up to 5.11% in the average FPR95 on the ImageNet benchmark.

## [Personalized Federated Learning towards Communication Efficiency, Robustness and Fairness](#)

- Shiyun Lin · Yuze Han · Xiang Li · Zhihua Zhang
- abstract@[open-review](#): Personalized Federated Learning faces many challenges such as expensive communication costs, training-time adversarial attacks, and performance unfairness across devices. Recent developments witness a trade-off between a reference model and local models to achieve personalization. We follow the avenue and propose a personalized FL method towards the three goals. When it is time to communicate, our method projects local models into a shared-and-fixed low-dimensional random subspace and uses infimal convolution to control the deviation between the reference model and projected local models. We theoretically show our method converges for strongly convex objectives with square regularizers and the convergence dependence on the projection dimension is mild. We also illustrate the benefits of robustness and fairness on a class of linear problems. Finally, we conduct a large number of experiments to show the empirical superiority of our method over several state-of-the-art methods on the three aspects.

## [MoVQ: Modulating Quantized Vectors for High-Fidelity Image Generation](#)

- Chuanxia Zheng · Tung-Long Vuong · Jianfei Cai · Dinh Phung
- abstract@[open-review](#): Although two-stage Vector Quantized (VQ) generative models allow for synthesizing high-fidelity and high-resolution images, their quantization operator encodes similar patches within an image into the same index, resulting in a repeated artifact for similar adjacent regions using existing decoder architectures. To address this issue, we propose to incorporate the spatially conditional normalization to modulate the quantized vectors so as to insert spatially variant information to the embedded index maps, encouraging the decoder to generate more photorealistic images. Moreover, we use multichannel quantization to increase the recombination capability of the discrete codes without increasing the cost of model and codebook. Additionally, to generate discrete tokens at the second stage, we adopt a Masked Generative Image Transformer (MaskGIT) to learn an underlying prior distribution in the compressed latent space, which is much faster than the conventional autoregressive model. Experiments on two benchmark datasets

demonstrate that our proposed modulated VQGAN is able to greatly improve the reconstructed image quality as well as provide high-fidelity image generation.

## [Learning Causally Invariant Representations for Out-of-Distribution Generalization on Graphs](#)

- Yongqiang Chen · Yonggang Zhang · Yatao Bian · Han Yang · MA Kaili · Binghui Xie · Tongliang Liu · Bo Han · James Cheng
- abstract@[open-review](#): Despite recent success in using the invariance principle for out-of-distribution (OOD) generalization on Euclidean data (e.g., images), studies on graph data are still limited. Different from images, the complex nature of graphs poses unique challenges to adopting the invariance principle. In particular, distribution shifts on graphs can appear in a variety of forms such as attributes and structures, making it difficult to identify the invariance. Moreover, domain or environment partitions, which are often required by OOD methods on Euclidean data, could be highly expensive to obtain for graphs. To bridge this gap, we propose a new framework, called Causality Inspired Invariant Graph LeArning (CIGA), to capture the invariance of graphs for guaranteed OOD generalization under various distribution shifts. Specifically, we characterize potential distribution shifts on graphs with causal models, concluding that OOD generalization on graphs is achievable when models focus only on subgraphs containing the most information about the causes of labels. Accordingly, we propose an information-theoretic objective to extract the desired subgraphs that maximally preserve the invariant intra-class information. Learning with these subgraphs is immune to distribution shifts. Extensive experiments on both synthetic and real-world datasets, including a challenging setting in AI-aided drug discovery, validate the superior OOD generalization ability of CIGA.

## [Graph Learning Assisted Multi-Objective Integer Programming](#)

- Yaxin Wu · Wen Song · Zhiguang Cao · Jie Zhang · Abhishek Gupta · Mingyan Lin
- abstract@[open-review](#): Objective-space decomposition algorithms (ODAs) are widely studied for solving multi-objective integer programs. However, they often encounter difficulties in handling scalarized problems, which could cause infeasibility or repetitive nondominated points and thus induce redundant runtime. To mitigate the issue, we present a graph neural network (GNN) based method to learn the reduction rule in the ODA. We formulate the algorithmic procedure of generic ODAs as a Markov decision process, and parameterize the policy (reduction rule) with a novel two-stage GNN to fuse information from variables, constraints and especially objectives for better state representation. We train our model with imitation learning and deploy it on a state-of-the-art ODA. Results show that our method significantly improves the solving efficiency of the ODA. The learned policy generalizes fairly well to larger problems or more objectives, and the proposed GNN outperforms existing ones for integer programming in terms of test and generalization accuracy.

## [Missing Data Imputation and Acquisition with Deep Hierarchical Models and Hamiltonian Monte Carlo](#)

- Ignacio Peis · Chao Ma · José Miguel Hernández-Lobato
- abstract@[open-review](#): Variational Autoencoders (VAEs) have recently been highly successful at imputing and acquiring heterogeneous missing data. However, within this specific application domain, existing VAE methods are restricted by using only one layer of latent variables and strictly Gaussian posterior approximations. To address these limitations, we present HH-VAEM, a Hierarchical VAE model for mixed-type incomplete data that uses Hamiltonian Monte Carlo with automatic hyper-parameter tuning for improved approximate inference. Our experiments show that HH-VAEM outperforms existing baselines in the tasks of missing data imputation and supervised learning with missing features. Finally, we also present a sampling-based approach for efficiently computing the information gain when missing features are to be acquired with HH-VAEM. Our experiments show that this sampling-based approach is superior to alternatives based on Gaussian approximations.

## [PolarMix: A General Data Augmentation Technique for LiDAR Point Clouds](#)

- Aoran Xiao · Jiaxing Huang · Dayan Guan · Kaiwen Cui · Shijian Lu · Ling Shao
- abstract@[open-review](#): LiDAR point clouds, which are usually scanned by rotating LiDAR sensors continuously, capture precise geometry of the surrounding environment and are crucial to many autonomous detection and navigation tasks. Though many 3D deep architectures have been developed, efficient collection and annotation of large amounts of point clouds remain one major challenge in the analytics and understanding of point cloud data. This paper presents PolarMix, a point cloud augmentation technique that is simple and generic but can mitigate the data constraint effectively across various perception tasks and scenarios. PolarMix enriches point cloud distributions and preserves point cloud fidelity via two cross-scan augmentation strategies that cut, edit, and mix point clouds along the scanning direction. The first is scene-level swapping which exchanges point cloud sectors of two LiDAR scans that are cut along the LiDAR scanning direction. The second is instance-level rotation and paste which crops point instances from one LiDAR scan, rotates them by multiple angles (to create multiple copies), and paste the rotated point instances into other scans. Extensive experiments show that PolarMix achieves superior performance consistently across different perception tasks and scenarios. In addition, it can work as plug-and-play for various 3D deep architectures and also performs well for unsupervised domain adaptation. Code will be available.

## [Exact Shape Correspondence via 2D graph convolution](#)

- Barakeel Fanseu Kamhoua · Lin Zhang · Yongqiang Chen · Han Yang · MA Kaili · Bo Han · Bo Li · James Cheng
- abstract@[open-review](#): For exact 3D shape correspondence (matching or alignment), i.e., the task of matching each point on a shape to its exact corresponding point on the other shape (or to be more specific, matching at geodesic error 0), most existing methods do not perform well due to two main problems. First, on nearly-isometric shapes (i.e., low noise levels), most existing methods use the eigen-vectors (eigen-functions) of the Laplace Beltrami Operator (LBO) or other shape descriptors to update an initialized correspondence which is not exact, leading to an accumulation of update errors. Thus, though the final correspondence may generally be smooth, it is generally inexact. Second, on non-isometric shapes (noisy shapes), existing methods are generally not robust to noise as they usually assume near-isometry. In addition, existing methods that attempt to address the non-isometric shape problem (e.g., GRAMPA) are generally computationally expensive and do not generalise to nearly-isometric shapes. To address these two problems, we propose a 2D graph convolution-based framework called 2D-GEM. 2D-GEM is robust to noise on non-isometric shapes and with a few additional constraints, it also addresses the errors in the update on nearly-isometric shapes. We demonstrate the effectiveness of 2D-GEM by achieving a high accuracy of 90.5% at geodesic error 0 on the non-isometric benchmark SHREC16, i.e., TOPKIDS (while being much faster than GRAMPA), and on nearly-isometric benchmarks by achieving a high accuracy of 92.5% on TOSCA and 84.9% on SCAPE at geodesic error 0.

## [Support Recovery in Sparse PCA with Incomplete Data](#)

- Hanbyul Lee · Qifan Song · Jean Honorio
- abstract@[open-review](#): We study a practical algorithm for sparse principal component analysis (PCA) of incomplete and noisy data. Our algorithm is based on the semidefinite program (SDP) relaxation of the non-convex  $\| \cdot \|_1$ -regularized PCA problem. We provide theoretical and experimental evidence that SDP enables us to exactly recover the true support of the sparse leading eigenvector of the unknown true matrix, despite only observing an incomplete (missing uniformly at random) and noisy version of it. We derive sufficient conditions for exact recovery, which involve matrix incoherence, the spectral gap between the largest and second-largest eigenvalues, the observation probability and the noise variance. We validate our theoretical results with incomplete synthetic data, and show encouraging and meaningful results on a gene expression dataset.

## [Divide and Contrast: Source-free Domain Adaptation via Adaptive Contrastive Learning](#)

- Ziyi Zhang · Weikai Chen · Hui Cheng · Zhen Li · Siyuan Li · Liang Lin · Guanbin Li
- abstract@[open-review](#): We investigate a practical domain adaptation task, called source-free domain adaptation (SFUDA), where the source pretrained model is adapted to the target domain without access to the source data. Existing techniques mainly leverage self-supervised pseudo-labeling to achieve class-wise global alignment [1] or rely on local structure extraction that encourages the feature consistency among neighborhoods [2]. While impressive progress has been made, both lines of methods have their own drawbacks — the “global” approach is sensitive to noisy labels while the “local” counterpart suffers from the source bias. In this paper, we present Divide and Contrast (DaC), a new paradigm for SFUDA that strives to connect the good ends of both worlds while bypassing their limitations. Based on the prediction confidence of the source model, DaC divides the target data into source-like and target-specific samples, where either group of samples is treated with tailored goals under an adaptive contrastive learning framework. Specifically, the source-like samples are utilized for learning global class clustering thanks to their relatively clean labels. The more noisy target-specific data are harnessed at the instance level for learning the intrinsic local structures. We further align the source-like domain with the target-specific samples using a memory bank-based Maximum Mean Discrepancy (MMD) loss to reduce the distribution mismatch. Extensive experiments on VisDA, Office-Home, and the more challenging DomainNet have verified the superior performance of DaC over current state-of-the-art approaches. The code is available at <https://github.com/ZyeZhang/DaC.git>.

## [Controllable Text Generation with Neurally-Decomposed Oracle](#)

- Tao Meng · Sidi Lu · Nanyun Peng · Kai-Wei Chang
- abstract@[open-review](#): We propose a general and efficient framework to control auto-regressive generation models with NeurALLY-Decomposed Oracle (NADO). Given a pre-trained base language model and a sequence-level boolean oracle function, we aim to decompose the oracle function into token-level guidance to steer the base model in text generation. Specifically, the token-level guidance is provided by NADO, a neural model trained with examples sampled from the base model, demanding no additional auxiliary labeled data. We present the close-form optimal solution to incorporate the decomposed token-level guidance into the base model for controllable generation. We further discuss how the neural approximation affects the quality of the solution. These experiments conducted on two different applications: (1) text generation with lexical constraints and (2) machine translation with formality control demonstrate that our framework efficiently guides the base model towards the given oracle while keeping high generation quality.

## [Lipschitz Bandits with Batched Feedback](#)

- Yasong Feng · zengfeng Huang · Tianyu Wang
- abstract@[open-review](#): In this paper, we study Lipschitz bandit problems with batched feedback, where the expected reward is Lipschitz and the reward observations are communicated to the player in batches. We introduce a novel landscape-aware algorithm, called Batched Lipschitz Narrowing (BLiN), that optimally solves this problem. Specifically, we show that for a  $T$ -step problem with Lipschitz reward of zooming dimension  $d_z$ , our algorithm achieves theoretically optimal (up to logarithmic factors) regret rate  $\tilde{O}(\sqrt{\frac{d_z+1}{d_z+2}}T)$  using only  $\Omega(\log \log T)$  batches. We also provide complexity analysis for this problem. Our theoretical lower bound implies that  $\Omega(\log \log T)$  batches are necessary for any algorithm to achieve the optimal regret. Thus, BLiN achieves optimal regret rate using minimal communication.

## [Learning Concept Credible Models for Mitigating Shortcuts](#)

- Jiaxuan Wang · Sarah Jabbour · Maggie Makar · Michael Sjoding · Jenna Wiens
- abstract@[open-review](#): During training, models can exploit spurious correlations as shortcuts, resulting in poor generalization performance when shortcuts do not persist. In this work, assuming access to a representation based on domain knowledge (i.e., known concepts) that is invariant to shortcuts, we aim to learn robust and accurate models from biased training data. In contrast to previous work, we do not rely solely on known concepts, but allow the model to also learn unknown concepts. We propose two approaches for mitigating shortcuts that incorporate domain knowledge, while accounting for potentially important yet unknown concepts. The first approach is two-staged. After fitting a model using known concepts, it accounts for the residual using unknown concepts. While flexible, we show that this approach is vulnerable when shortcuts are correlated with the unknown concepts. This limitation is addressed by our second approach that extends a recently proposed regularization penalty. Applied to two real-world datasets, we demonstrate that both approaches can successfully mitigate shortcut learning.

## [Physical Design using Differentiable Learned Simulators](#)

- Kelsey Allen · Tatiana Lopez-Guevara · Kimberly Stachenfeld · Alvaro Sanchez Gonzalez · Peter Battaglia · Jessica Hamrick · Tobias Pfaff
- abstract@[open-review](#): Designing physical artifacts that serve a purpose---such as tools and other functional structures---is central to engineering as well as everyday human behavior. Though automating design using machine learning has tremendous promise, existing methods are often limited by the task-dependent distributions they were exposed to during training. Here we showcase a task-agnostic approach to inverse design, by combining general-purpose graph network simulators with gradient-based design optimization. This constitutes a simple, fast, and reusable approach that solves high-dimensional problems with complex physical dynamics, including designing surfaces and tools to manipulate fluid flows and optimizing the shape of an airfoil to minimize drag. This framework produces high-quality designs by propagating gradients through trajectories of hundreds of steps, even when using models that were pre-trained for single-step predictions on data substantially different from the design tasks. In our fluid manipulation tasks, the resulting designs outperformed those found by sampling-based optimization techniques. In airfoil design, they matched the quality of those obtained with a specialized solver. Our results suggest that despite some remaining challenges, machine learning-based simulators are maturing to the point where they can support general-purpose design optimization across a variety of fluid-structure interaction domains.

## [Positively Weighted Kernel Quadrature via Subsampling](#)

- Satoshi Hayakawa · Harald Oberhauser · Terry Lyons
- abstract@[open-review](#): We study kernel quadrature rules with convex weights. Our approach combines the spectral properties of the kernel with recombination results about point measures. This results in effective algorithms that construct convex quadrature rules using only access to i.i.d. samples from the underlying measure and evaluation of the kernel and that result in a small worst-case error. In addition to our theoretical results and the benefits resulting from convex weights, our experiments indicate that this construction can compete with the optimal bounds in well-known examples.

## [Consistency of Constrained Spectral Clustering under Graph Induced Fair Planted Partitions](#)

- Shubham Gupta · Ambedkar Dukkipati
- abstract@[open-review](#): Spectral clustering is popular among practitioners and theoreticians alike. While performance guarantees for spectral clustering are well understood, recent studies have focused on enforcing fairness" in clusters, requiring them to be balanced" with respect to a categorical sensitive node attribute (e.g. the race distribution in clusters must match the race distribution in the population). In this paper, we consider a setting where sensitive attributes indirectly manifest in an auxiliary representation graph rather than being directly observed. This graph specifies node pairs that can represent each other with respect to sensitive attributes and is observed in addition to the usual similarity graph. Our goal is to find clusters in the similarity graph while respecting a new individual-level fairness constraint encoded by the representation graph. We develop variants of unnormalized and normalized spectral clustering for this task and analyze their performance under a fair planted partition model induced by the representation graph. This model uses both the cluster membership of the nodes and the structure of the representation graph to generate random similarity graphs. To the

best of our knowledge, these are the first consistency results for constrained spectral clustering under an individual-level fairness constraint. Numerical results corroborate our theoretical findings.

## [Hyper-Representations as Generative Models: Sampling Unseen Neural Network Weights](#)

- Konstantin Schörholt · Boris Knyazev · Xavier Giro-i-Nieto · Damian Borth
- abstract@[open-review](#): Learning representations of neural network weights given a model zoo is an emerging and challenging area with many potential applications from model inspection, to neural architecture search or knowledge distillation. Recently, an autoencoder trained on a model zoo was able to learn a hyper-representation, which captures intrinsic and extrinsic properties of the models in the zoo. In this work, we extend hyper-representations for generative use to sample new model weights. We propose layer-wise loss normalization which we demonstrate is key to generate high-performing models and several sampling methods based on the topology of hyper-representations. The models generated using our methods are diverse, performant and capable to outperform strong baselines as evaluated on several downstream tasks: initialization, ensemble sampling and transfer learning. Our results indicate the potential of knowledge aggregation from model zoos to new models via hyper-representations thereby paving the avenue for novel research directions.

## [Learning Neural Set Functions Under the Optimal Subset Oracle](#)

- Zijing Ou · Tingyang Xu · Qinliang Su · Yingzhen Li · Peilin Zhao · Yatao Bian
- abstract@[open-review](#): Learning set functions becomes increasingly more important in many applications like product recommendation and compound selection in AI-aided drug discovery. The majority of existing works study methodologies of set function learning under the function value oracle, which, however, requires expensive supervision signals. This renders it impractical for applications with only weak supervisions under the Optimal Subset (OS) oracle, the study of which is surprisingly overlooked. In this work, we present a principled yet practical maximum likelihood learning framework, termed as EquivSet, that simultaneously meets the following desiderata of learning set functions under the OS oracle: i) permutation invariance of the set mass function being modeled; ii) permission of varying ground set; iii) minimum prior and iv) scalability. The main components of our framework involve: an energy-based treatment of the set mass function, DeepSet-style architectures to handle permutation invariance, mean-field variational inference, and its amortized variants. Thanks to the delicate combination of these advanced architectures, empirical studies on three real-world applications (including Amazon product recommendation, set anomaly detection, and compound selection for virtual screening) demonstrate that EquivSet outperforms the baselines by a large margin.

## [An Analytical Theory of Curriculum Learning in Teacher-Student Networks](#)

- Luca Saglietti · Stefano Mannelli · Andrew Saxe
- abstract@[open-review](#): In animals and humans, curriculum learning---presenting data in a curated order---is critical to rapid learning and effective pedagogy. A long history of experiments has demonstrated the impact of curricula in a variety of animals but, despite its ubiquitous presence, a theoretical understanding of the phenomenon is still lacking. Surprisingly, in contrast to animal learning, curricula strategies are not widely used in machine learning and recent simulation studies reach the conclusion that curricula are moderately effective or ineffective in most cases. This stark difference in the importance of curriculum raises a fundamental theoretical question: when and why does curriculum learning help? In this work, we analyse a prototypical neural network model of curriculum learning in the high-dimensional limit, employing statistical physics methods. We study a task in which a sparse set of informative features are embedded amidst a large set of noisy features. We analytically derive average learning trajectories for simple neural networks on this task, which establish a clear speed benefit for curriculum learning in the online setting. However, when training experiences can be stored and replayed (for instance, during sleep), the advantage of curriculum in standard neural networks disappears, in line with observations from the deep learning literature. Inspired by synaptic consolidation techniques developed to combat catastrophic forgetting, we investigate whether consolidating synapses at curriculum change points can boost the benefits of curricula. We derive generalisation performance as a function of consolidation strength (implemented as a Gaussian prior connecting learning phases), and show that this consolidation mechanism can yield a large improvement in test performance. Our reduced analytical descriptions help reconcile apparently conflicting empirical results, trace regimes where curriculum learning yields the largest gains, and provide experimentally-accessible predictions for the impact of task parameters on curriculum benefits. More broadly, our results suggest that fully exploiting a curriculum may require explicit consolidation at curriculum boundaries.

## [When to intervene: learning optimal intervention policies for critical events](#)

- Niranjan Damera Venkata · Chiranjib Bhattacharyya
- abstract@[open-review](#): Providing timely interventions before the onset of a critical event, such as a system failure is an important problem in many industrial settings. Before the onset of the critical event, usually, systems exhibit behavioral changes which often manifest as stochastic co-variate observations which may be leveraged to trigger intervention. In this paper, for the first time, we formulate the problem of finding an optimally timed intervention (OTI) policy as minimizing the expected residual time to event, subject to a constraint on the probability of missing the event. Existing machine learning approaches to intervention on critical events focus on predicting event occurrence within a pre-defined window (a classification problem) or predicting time-to-event (a regression problem). Interventions are then triggered by setting model thresholds. These are heuristic-driven, lacking guarantees regarding optimality. To model the evolution of system behavior, we introduce the concept of a hazard rate process. We show that the OTI problem is equivalent to an optimal stopping problem on the associated hazard rate process. This key link has not been explored in literature. Under Markovian assumptions on the hazard rate process, we show that an OTI policy at any time is completely determined by the conditional hazard rate function at that time. We proceed to analytically characterize the optimal stopping policy, enabling practical survival-model-based critical event intervention algorithms. Further, we show that our framework includes, as a special case, the important class of neural hazard rate processes generated by recurrent neural networks (RNNs). To model neural hazard rate processes we propose a dynamic deep recurrent survival analysis (DDRSA) architecture, introducing an RNN encoder into the static DRSA setting. Finally, we demonstrate RNN-based OTI policies with practical experiments and show that they outperform popular intervention methods.

## [Collaborative Decision Making Using Action Suggestions](#)

- Dylan Asmar · Mykel J Kochenderfer
- abstract@[open-review](#): The level of autonomy is increasing in systems spanning multiple domains, but these systems still experience failures. One way to mitigate the risk of failures is to integrate human oversight of the autonomous systems and rely on the human to take control when the autonomy fails. In this work, we formulate a method of collaborative decision making through action suggestions that improves action selection without taking control of the system. Our approach uses each suggestion efficiently by incorporating the implicit information shared through suggestions to modify the agent's belief and achieves better performance with fewer suggestions than naively following the suggested actions. We assume collaborative agents share the same objective and communicate through valid actions. By assuming the suggested action is dependent only on the state, we can incorporate the suggested action as an independent observation of the environment. The assumption of a collaborative environment enables us to use the agent's policy to estimate the distribution over action suggestions. We propose two methods that use suggested actions and demonstrate the approach through simulated experiments. The proposed methodology results in increased performance while also being robust to suboptimal suggestions.

## [S2P: State-conditioned Image Synthesis for Data Augmentation in Offline Reinforcement Learning](#)

- Daesol Cho · Dongseok Shim · H. Jin Kim

- abstract@[open-review](#): Offline reinforcement learning (Offline RL) suffers from the innate distributional shift as it cannot interact with the physical environment during training. To alleviate such limitation, state-based offline RL leverages a learned dynamics model from the logged experience and augments the predicted state transition to extend the data distribution. For exploiting such benefit also on the image-based RL, we firstly propose a generative model, S2P (State2Pixel), which synthesizes the raw pixel of the agent from its corresponding state. It enables bridging the gap between the state and the image domain in RL algorithms, and virtually exploring unseen image distribution via model-based transition in the state space. Through experiments, we confirm that our S2P-based image synthesis not only improves the image-based offline RL performance but also shows powerful generalization capability on unseen tasks.

## [Chromatic Correlation Clustering, Revisited](#)

- Qing Xiu · Kai Han · Jing Tang · Shuang Cui · He Huang
- abstract@[open-review](#): Chromatic Correlation Clustering (CCC) (introduced by Bonchi et al. [6]) is a natural generalization of the celebrated Correlation Clustering (CC) problem, introduced by Bonchi et al. [6]. It models objects with categorical pairwise relationships by an edge-colored graph, and has many applications in data mining, social networks and bioinformatics. We show that there exists a \$2.5\\$-approximation to the CCC problem based on a Linear Programming (LP) approach, thus improving the best-known approximation ratio of 3 achieved by Klodt et al. [21] . We also present an efficient heuristic algorithm for CCC leveraging a greedy clustering strategy, and conduct extensive experiments to demonstrate the effectiveness and efficiency of our proposed algorithm.

## [Procedural Image Programs for Representation Learning](#)

- Manel Baradad · Richard Chen · Jonas Wulff · Tongzhou Wang · Rogerio Feris · Antonio Torralba · Phillip Isola
- abstract@[open-review](#): Learning image representations using synthetic data allows training neural networks without some of the concerns associated with real images, such as privacy and bias. Existing work focuses on a handful of generative processes which are hard to integrate together to scale up. To overcome this, we propose training with a large dataset of twenty-one thousand programs, each one generating a diverse set of synthetic images. These programs are short code snippets, which are easy to modify and fast to execute using OpenGL. The proposed dataset can be used for both supervised and unsupervised representation learning, and reduces the gap between pre-training with real and procedurally generated images by 38%.

## [Text-Adaptive Multiple Visual Prototype Matching for Video-Text Retrieval](#)

- Chengzhi Lin · Ancong Wu · Junwei Liang · Jun Zhang · Wenhong Ge · Wei-Shi Zheng · Chunhua Shen
- abstract@[open-review](#): Cross-modal retrieval between videos and texts has gained increasing interest because of the rapid emergence of videos on the web. Generally, a video contains rich instance and event information and the query text only describes a part of the information. Thus, a video can have multiple different text descriptions and queries. We call it the Video-Text Correspondence Ambiguity problem. Current techniques mostly concentrate on mining local or multi-level alignment between contents of video and text (e.g., object to entity and action to verb). It is difficult for these methods to alleviate video-text correspondence ambiguity by describing a video using only one feature, which is required to be matched with multiple different text features at the same time. To address this problem, we propose a Text-Adaptive Multiple Visual Prototype Matching Model. It automatically captures multiple prototypes to describe a video by adaptive aggregation on video token features. Given a query text, the similarity is determined by the most similar prototype to find correspondence in the video, which is called text-adaptive matching. To learn diverse prototypes for representing the rich information in videos, we propose a variance loss to encourage different prototypes to attend to different contents of the video. Our method outperforms state-of-the-art methods on four public video retrieval datasets.

## [Deciding What to Model: Value-Equivalent Sampling for Reinforcement Learning](#)

- Dilip Arumugam · Benjamin Van Roy
- abstract@[open-review](#): The quintessential model-based reinforcement-learning agent iteratively refines its estimates or prior beliefs about the true underlying model of the environment. Recent empirical successes in model-based reinforcement learning with function approximation, however, eschew the true model in favor of a surrogate that, while ignoring various facets of the environment, still facilitates effective planning over behaviors. Recently formalized as the value equivalence principle, this algorithmic technique is perhaps unavoidable as real-world reinforcement learning demands consideration of a simple, computationally-bounded agent interacting with an overwhelmingly complex environment, whose underlying dynamics likely exceed the agent's capacity for representation. In this work, we consider the scenario where agent limitations may entirely preclude identifying an exactly value-equivalent model, immediately giving rise to a trade-off between identifying a model that is simple enough to learn while only incurring bounded sub-optimality. To address this problem, we introduce an algorithm that, using rate-distortion theory, iteratively computes an approximately-value-equivalent, lossy compression of the environment which an agent may feasibly target in lieu of the true model. We prove an information-theoretic, Bayesian regret bound for our algorithm that holds for any finite-horizon, episodic sequential decision-making problem. Crucially, our regret bound can be expressed in one of two possible forms, providing a performance guarantee for finding either the simplest model that achieves a desired sub-optimality gap or, alternatively, the best model given a limit on agent capacity.

## [Mind Reader: Reconstructing complex images from brain activities](#)

- Sikun Lin · Thomas Sprague · Ambuj K Singh
- abstract@[open-review](#): Understanding how the brain encodes external stimuli and how these stimuli can be decoded from the measured brain activities are long-standing and challenging questions in neuroscience. In this paper, we focus on reconstructing the complex image stimuli from fMRI (functional magnetic resonance imaging) signals. Unlike previous works that reconstruct images with single objects or simple shapes, our work aims to reconstruct image stimuli that are rich in semantics, closer to everyday scenes, and can reveal more perspectives. However, data scarcity of fMRI datasets is the main obstacle to applying state-of-the-art deep learning models to this problem. We find that incorporating an additional text modality is beneficial for the reconstruction problem compared to directly translating brain signals to images. Therefore, the modalities involved in our method are: (i) voxel-level fMRI signals, (ii) observed images that trigger the brain signals, and (iii) textual description of the images. To further address data scarcity, we leverage an aligned vision-language latent space pre-trained on massive datasets. Instead of training models from scratch to find a latent space shared by the three modalities, we encode fMRI signals into this pre-aligned latent space. Then, conditioned on embeddings in this space, we reconstruct images with a generative model. The reconstructed images from our pipeline balance both naturalness and fidelity: they are photo-realistic and capture the ground truth image contents well.

## [Transform Once: Efficient Operator Learning in Frequency Domain](#)

- Michael Poli · Stefano Massaroli · Federico Berto · Jinkyoo Park · Tri Dao · Christopher RÃ© · Stefano Ermon
- abstract@[open-review](#): Spectrum analysis provides one of the most effective paradigms for information-preserving dimensionality reduction in data: often, a simple description of naturally occurring signals can be obtained via few terms of periodic basis functions. Neural operators designed for frequency domain learning -- frequency domain models (FDMs) -- are based on complex-valued transforms i.e. Fourier Transforms (FT), and layers that perform computation on the spectrum and input data separately. This design introduces considerable computational overhead: for each layer, a forward and inverse FT. Instead, this work introduces a blueprint for frequency domain learning through a single transform: transform once (T1). To enable efficient, direct learning in the frequency domain we develop a variance preserving weight initialization scheme and investigate various choices of transforms. Our results noticeably streamline the design process of FDMs, pruning redundant transforms, and leading to speedups of 3x to 10x that

increase with data resolution and model size. We perform extensive experiments on learning to solve partial differential equations, including incompressible Navier-Stokes, turbulent flows around airfoils, and high-resolution video of smoke dynamics. T1 models improve on the test performance of SOTA FDMs while requiring significantly less computation, with over 20% reduction in predictive error across tasks.

## [Holomorphic Equilibrium Propagation Computes Exact Gradients Through Finite Size Oscillations](#)

- Axel Laborieux Å· Friedemann Zenke
- abstract@[open-review](#): Equilibrium propagation (EP) is an alternative to backpropagation (BP) that allows the training of deep neural networks with local learning rules. It thus provides a compelling framework for training neuromorphic systems and understanding learning in neurobiology. However, EP requires infinitesimal teaching signals, thereby limiting its applicability to noisy physical systems. Moreover, the algorithm requires separate temporal phases and has not been applied to large-scale problems. Here we address these issues by extending EP to holomorphic networks. We show analytically that this extension naturally leads to exact gradients for finite-amplitude teaching signals. Importantly, the gradient can be computed as the first Fourier coefficient from finite neuronal activity oscillations in continuous time without requiring separate phases. Further, we demonstrate in numerical simulations that our approach permits robust estimation of gradients in the presence of noise and that deeper models benefit from the finite teaching signals. Finally, we establish the first benchmark for EP on the ImageNet \$32 \times 32\$ dataset and show that it matches the performance of an equivalent network trained with BP. Our work provides analytical insights that enable scaling EP to large-scale problems and establishes a formal framework for how oscillations could support learning in biological and neuromorphic systems.

## [Reinforced Genetic Algorithm for Structure-based Drug Design](#)

- Tianfan Fu Å· Wenhao Gao Å· Connor Coley Å· Jimeng Sun
- abstract@[open-review](#): Structure-based drug design (SBDD) aims to discover drug candidates by finding molecules (ligands) that bind tightly to a disease-related protein (targets), which is the primary approach to computer-aided drug discovery. Recently, applying deep generative models for three-dimensional (3D) molecular design conditioned on protein pockets to solve SBDD has attracted much attention, but their formulation as probabilistic modeling often leads to unsatisfactory optimization performance. On the other hand, traditional combinatorial optimization methods such as genetic algorithms (GA) have demonstrated state-of-the-art performance in various molecular optimization tasks. However, they do not utilize protein target structure to inform design steps but rely on a random-walk-like exploration, which leads to unstable performance and no knowledge transfer between different tasks despite the similar binding physics. To achieve a more stable and efficient SBDD, we propose Reinforced Genetic Algorithm (RGA) that uses neural models to prioritize the profitable design steps and suppress random-walk behavior. The neural models take the 3D structure of the targets and ligands as inputs and are pre-trained using native complex structures to utilize the knowledge of the shared binding physics from different targets and then fine-tuned during optimization. We conduct thorough empirical studies on optimizing binding affinity to various disease targets and show that RGA outperforms the baselines in terms of docking scores and is more robust to random initializations. The ablation study also indicates that the training on different targets helps improve the performance by leveraging the shared underlying physics of the binding processes.

## [TA-MoE: Topology-Aware Large Scale Mixture-of-Expert Training](#)

- Chang Chen Å· Min Li Å· Zhihua Wu Å· Dianhai Yu Å· Chao Yang
- abstract@[open-review](#): Sparsely gated Mixture-of-Expert (MoE) has demonstrated its effectiveness in scaling up deep neural networks to an extreme scale. Despite that numerous efforts have been made to improve the performance of MoE from the model design or system optimization perspective, existing MoE dispatch patterns are still not able to fully exploit the underlying heterogeneous network environments. In this paper, we propose TA-MoE, a topology-aware routing strategy for large-scale MoE training, from a model-system co-design perspective, which can dynamically adjust the MoE dispatch pattern according to the network topology. Based on communication modeling, we abstract the dispatch problem into an optimization objective and obtain the approximate dispatch pattern under different topologies. On top of that, we design a topology-aware auxiliary loss, which can adaptively route the data to fit in the underlying topology without sacrificing the model accuracy. Experiments show that TA-MoE can substantially outperform its counterparts on various hardware and model configurations, with roughly 1.01x-1.61x, 1.01x-4.77x, 1.25x-1.54x improvements over the popular DeepSpeed-MoE, FastMoE and FasterMoE systems.

## [Learning Dynamical Systems via Koopman Operator Regression in Reproducing Kernel Hilbert Spaces](#)

- Pietro Novelli Å· Vladimir Kostic Å· Massimiliano Pontil Å· Andreas Maurer Å· Carlo Ciliberto Å· Lorenzo Rosasco
- abstract@[open-review](#): We study a class of dynamical systems modelled as stationary Markov chains that admit an invariant distribution via the corresponding transfer or Koopman operator. While data-driven algorithms to reconstruct such operators are well known, their relationship with statistical learning is largely unexplored. We formalize a framework to learn the Koopman operator from finite data trajectories of the dynamical system. We consider the restriction of this operator to a reproducing kernel Hilbert space and introduce a notion of risk, from which different estimators naturally arise. We link the risk with the estimation of the spectral decomposition of the Koopman operator. These observations motivate a reduced-rank operator regression (RRR) estimator. We derive learning bounds for the proposed estimator, holding both in i.i.d and non i.i.d. settings, the latter in terms of mixing coefficients. Our results suggest RRR might be beneficial over other widely used estimators as confirmed in numerical experiments both for forecasting and mode decomposition.

## [Cooperative Distribution Alignment via JSD Upper Bound](#)

- Wonwoong Cho Å· ZIYU GONG Å· David Inouye
- abstract@[open-review](#): Unsupervised distribution alignment estimates a transformation that maps two or more source distributions to a shared aligned distribution given only samples from each distribution. This task has many applications including generative modeling, unsupervised domain adaptation, and socially aware learning. Most prior works use adversarial learning (i.e., min-max optimization), which can be challenging to optimize and evaluate. A few recent works explore non-adversarial flow-based (i.e., invertible) approaches, but they lack a unified perspective and are limited in efficiently aligning multiple distributions. Therefore, we propose to unify and generalize previous flow-based approaches under a single non-adversarial framework, which we prove is equivalent to minimizing an upper bound on the Jensen-Shannon Divergence (JSD). Importantly, our problem reduces to a min-min, i.e., cooperative, problem and can provide a natural evaluation metric for unsupervised distribution alignment. We present empirical results of our framework on both simulated and real-world datasets to demonstrate the benefits of our approach.

## [Semi-Supervised Learning with Decision Trees: Graph Laplacian Tree Alternating Optimization](#)

- Arman Zharmagambetov Å· Miguel A. Carreira-Perpinan
- abstract@[open-review](#): Semi-supervised learning seeks to learn a machine learning model when only a small amount of the available data is labeled. The most widespread approach uses a graph prior, which encourages similar instances to have similar predictions. This has been very successful with models ranging from kernel machines to neural networks, but has remained inapplicable to decision trees, for which the optimization problem is much harder. We solve this based on a reformulation of the problem which requires iteratively solving two simpler problems: a supervised tree learning problem, which can be solved by the Tree Alternating Optimization algorithm; and a label smoothing problem, which can be solved through a sparse linear system. The algorithm is scalable and highly effective even with very few labeled instances, and makes it possible to learn accurate, interpretable models based on decision trees in such situations.

## [Multi-block Min-max Bilevel Optimization with Applications in Multi-task Deep AUC Maximization](#)

- Quanqi Hu · YONGJIAN ZHONG · Tianbao Yang
- abstract@[open-review](#): In this paper, we study multi-block min-max bilevel optimization problems, where the upper level is non-convex strongly-concave minimax objective and the lower level is a strongly convex objective, and there are multiple blocks of dual variables and lower level problems. Due to the intertwined multi-block min-max bilevel structure, the computational cost at each iteration could be prohibitively high, especially with a large number of blocks. To tackle this challenge, we present a single-loop randomized stochastic algorithm, which requires updates for only a constant number of blocks at each iteration. Under some mild assumptions on the problem, we establish its sample complexity of  $\mathcal{O}(1/\epsilon^4)$  for finding an  $\epsilon$ -stationary point. This matches the optimal complexity order for solving stochastic nonconvex optimization under a general unbiased stochastic oracle model. Moreover, we provide two applications of the proposed method in multi-task deep AUC (area under ROC curve) maximization. Experimental results validate our theory and demonstrate the effectiveness of our method.

## [CLEAR: Generative Counterfactual Explanations on Graphs](#)

- Jing Ma · Ruocheng Guo · Saumitra Mishra · Aidong Zhang · Jundong Li
- abstract@[open-review](#): Counterfactual explanations promote explainability in machine learning models by answering the question "how should the input instance be altered to obtain a desired predicted label?". The comparison of this instance before and after perturbation can enhance human interpretation. Most existing studies on counterfactual explanations are limited in tabular data or image data. In this paper, we study the problem of counterfactual explanation generation on graphs. A few studies have explored to generate counterfactual explanations on graphs, but many challenges of this problem are still not well-addressed: 1) optimizing in the discrete and disorganized space of graphs; 2) generalizing on unseen graphs; 3) maintaining the causality in the generated counterfactuals without prior knowledge of the causal model. To tackle these challenges, we propose a novel framework CLEAR which aims to generate counterfactual explanations on graphs for graph-level prediction models. Specifically, CLEAR leverages a graph variational autoencoder based mechanism to facilitate its optimization and generalization, and promotes causality by leveraging an auxiliary variable to better identify the causal model. Extensive experiments on both synthetic and real-world graphs validate the superiority of CLEAR over state-of-the-art counterfactual explanation methods on graphs in different aspects.

## [VectorAdam for Rotation-Equivariant Geometry Optimization](#)

- Selena Zihan Ling · Nicholas Sharp · Alec Jacobson
- abstract@[open-review](#): The Adam optimization algorithm has proven remarkably effective for optimization problems across machine learning and even traditional tasks in geometry processing. At the same time, the development of equivariant methods, which preserve their output under the action of rotation or some other transformation, has proven to be important for geometry problems across these domains. In this work, we observe that naively applying Adam to optimize vector-valued data is not rotation-equivariant, due to per-coordinate moment updates, and in fact this leads to significant artifacts and biases in practice. We propose to resolve this deficiency with VectorAdam, a simple modification which makes Adam rotation-equivariant by accounting for the vector structure of optimization variables. We demonstrate this approach on common geometric optimization problems in traditional geometry processing and machine learning, showing that equivariant VectorAdam resolves the artifacts and biases of traditional Adam when applied to vector-valued data, with equivalent or even improved rates of convergence.

## [Flexible Diffusion Modeling of Long Videos](#)

- William Harvey · Saeid Naderiparizi · Vaden Masrani · Christian Weilbach · Frank Wood
- abstract@[open-review](#): We present a framework for video modeling based on denoising diffusion probabilistic models that produces long-duration video completions in a variety of realistic environments. We introduce a generative model that can at test-time sample any arbitrary subset of video frames conditioned on any other subset and present an architecture adapted for this purpose. Doing so allows us to efficiently compare and optimize a variety of schedules for the order in which frames in a long video are sampled and use selective sparse and long-range conditioning on previously sampled frames. We demonstrate improved video modeling over prior work on a number of datasets and sample temporally coherent videos over 25 minutes in length. We additionally release a new video modeling dataset and semantically meaningful metrics based on videos generated in the CARLA self-driving car simulator.

## [NSNet: A General Neural Probabilistic Framework for Satisfiability Problems](#)

- Zhaoyu Li · Xujie Si
- abstract@[open-review](#): We present the Neural Satisfiability Network (NSNet), a general neural framework that models satisfiability problems as probabilistic inference and meanwhile exhibits proper explainability. Inspired by the Belief Propagation (BP), NSNet uses a novel graph neural network (GNN) to parameterize BP in the latent space, where its hidden representations maintain the same probabilistic interpretation as BP. NSNet can be flexibly configured to solve both SAT and #SAT problems by supplying different learning objectives. For SAT, instead of directly predicting a satisfying assignment, NSNet performs marginal inference among all satisfying solutions, which we empirically find is more feasible for neural networks to learn. With the estimated marginals, a satisfying assignment can be efficiently generated by executing a stochastic local search. For #SAT, NSNet performs approximate model counting by learning the Bethe approximation of the partition function. Our evaluations show that NSNet achieves competitive results in terms of inference accuracy and time efficiency on multiple SAT and #SAT benchmarks.

## [BayesPCN: A Continually Learnable Predictive Coding Associative Memory](#)

- Jinsoo Yoo · Frank Wood
- abstract@[open-review](#): Associative memory plays an important role in human intelligence and its mechanisms have been linked to attention in machine learning. While the machine learning community's interest in associative memories has recently been rekindled, most work has focused on memory recall (\$read\$) over memory learning (\$write\$). In this paper, we present BayesPCN, a hierarchical associative memory capable of performing continual one-shot memory writes without meta-learning. Moreover, BayesPCN is able to gradually forget past observations (\$forget\$) to free its memory. Experiments show that BayesPCN can recall corrupted i.i.d. high-dimensional data observed hundreds of "timesteps" ago without a significant drop in recall ability compared to the state-of-the-art offline-learned associative memory models.

## [Non-monotonic Resource Utilization in the Bandits with Knapsacks Problem](#)

- Raunak Kumar · Robert Kleinberg
- abstract@[open-review](#): Bandits with knapsacks (BwK) is an influential model of sequential decision-making under uncertainty that incorporates resource consumption constraints. In each round, the decision-maker observes an outcome consisting of a reward and a vector of nonnegative resource consumptions, and the budget of each resource is decremented by its consumption. In this paper we introduce a natural generalization of the stochastic BwK problem that allows non-monotonic resource utilization. In each round, the decision-maker observes an outcome consisting of a reward and a vector of resource drifts that can be positive, negative or zero, and the budget of each resource is incremented by its drift. Our main result is a Markov decision process (MDP) policy that has constant regret against a linear programming (LP) relaxation when the decision-maker knows the true outcome distributions.

We build upon this to develop a learning algorithm that has logarithmic regret against the same LP relaxation when the decision-maker does not know the true outcome distributions. We also present a reduction from BwK to our model that shows our regret bound matches existing results.

## [TreeMoCo: Contrastive Neuron Morphology Representation Learning](#)

- Hanbo Chen · Jiawei Yang · Daniel Iascone · Lijuan Liu · Lei He · Hanchuan Peng · Jianhua Yao
- abstract@[open-review](#): Morphology of neuron trees is a key indicator to delineate neuronal cell-types, analyze brain development process, and evaluate pathological changes in neurological diseases. Traditional analysis mostly relies on heuristic features and visual inspections. A quantitative, informative, and comprehensive representation of neuron morphology is largely absent but desired. To fill this gap, in this work, we adopt a Tree-LSTM network to encode neuron morphology and introduce a self-supervised learning framework named TreeMoCo to learn features without the need for labels. We test TreeMoCo on 2403 high-quality 3D neuron reconstructions of mouse brains from three different public resources. Our results show that TreeMoCo is effective in both classifying major brain cell-types and identifying sub-types. To our best knowledge, TreeMoCo is the very first to explore learning the representation of neuron tree morphology with contrastive learning. It has a great potential to shed new light on quantitative neuron morphology analysis. Code is available at <https://github.com/TencentAILabHealthcare/NeuronRepresentation>.

## [Measuring Model Inversion Defences in Edge-Cloud Collaborative Inference Systems](#)

- Mengda Yang · Juan Wang · Hongxin Hu · Ao Ren · Ziang Li · Xiaoyang Xu · Wenzhe Yi
- abstract@[open-review](#): The edge-cloud collaborative inference systems are designed to speed up the prediction processes in edge-cloud scenarios, where the local devices and the cloud system work together to run a complex deep learning model. However, those edge-cloud collaborative inference systems are vulnerable to emerging Model Inversion (MI) attacks, where malicious cloud service providers are able to recover the edge-side users' private data. To defend against such attacks, several countermeasures have been recently introduced. Unfortunately, little is known about the robustness of those defence countermeasures. In this paper, we take the first step towards measuring the robustness of those state-of-the-art defence countermeasures with respect to MI attacks. We show the trade-offs of privacy and utility of these countermeasures and propose a novel anti-defence method called Sensitive Feature Distillation (SFD) to restore sensitive information from the protected feature representations. Our experiments show that SFD is able to break through defence mechanisms in model partitioning scenarios, demonstrating the inadequacy of existing defence mechanisms as a privacy-preserving technique against MI attacks. We hope our findings encourage researchers to pursue more robust defence mechanisms against MI attacks for edge-cloud collaborative inference systems.

## [Uncalibrated Models Can Improve Human-AI Collaboration](#)

- Kailas Vodrahalli · Tobias Gerstenberg · James Zou
- abstract@[open-review](#): In many practical applications of AI, an AI model is used as a decision aid for human users. The AI provides advice that a human (sometimes) incorporates into their decision-making process. The AI advice is often presented with some measure of "confidence" that the human can use to calibrate how much they depend on or trust the advice. In this paper, we present an initial exploration that suggests showing AI models as more confident than they actually are, even when the original AI is well-calibrated, can improve human-AI performance (measured as the accuracy and confidence of the human's final prediction after seeing the AI advice). We first train a model to predict human incorporation of AI advice using data from thousands of human interactions. This enables us to explicitly estimate how to transform the AI's prediction confidence, making the AI uncalibrated, in order to improve the final human prediction. We empirically validate our results across four different tasks---dealing with images, text and tabular data---involving hundreds of human participants. We further support our findings with simulation analysis. Our findings suggest the importance of jointly optimizing the human-AI system as opposed to the standard paradigm of optimizing the AI model alone.

## [Compositional Generalization in Unsupervised Representation Learning: From Disentanglement to Emergent Language](#)

- Zhenlin Xu · Marc Niethammer · Colin Raffel
- abstract@[open-review](#): Deep learning models struggle with compositional generalization, i.e.\ the ability to recognize or generate novel combinations of observed elementary concepts. In hopes of enabling compositional generalization, various unsupervised learning algorithms have been proposed with inductive biases that aim to induce compositional structure in learned representations (e.g.\ disentangled representation and emergent language learning). In this work, we evaluate these unsupervised learning algorithms in terms of how well they enable \textit{compositional generalization}. Specifically, our evaluation protocol focuses on whether or not it is easy to train a simple model on top of the learned representation that generalizes to new combinations of compositional factors. We systematically study three unsupervised representation learning algorithms -- \$\beta\$-VAE, \$\beta\$-TCVAE, and emergent language (EL) autoencoders -- on two datasets that allow directly testing compositional generalization. We find that directly using the bottleneck representation with simple models and few labels may lead to worse generalization than using representations from layers before or after the learned representation itself. In addition, we find that the previously proposed metrics for evaluating the levels of compositionality are not correlated with actual compositional generalization in our framework. Surprisingly, we find that increasing pressure to produce a disentangled representation (e.g.\ increasing \$\beta\$ in the \$\beta\$-VAE) produces representations with \textit{worse} generalization, while representations from EL models show strong compositional generalization. Motivated by this observation, we further investigate the advantages of using EL to induce compositional structure in unsupervised representation learning, finding that it shows consistently stronger generalization than disentanglement models, especially when using less unlabeled data for unsupervised learning and less labels for downstream tasks. Taken together, our results shed new light onto the compositional generalization behavior of different unsupervised learning algorithms with a new setting to rigorously test this behavior, and suggest the potential benefits of developing EL learning algorithms for more generalizable representations.

## [Towards Optimal Communication Complexity in Distributed Non-Convex Optimization](#)

- Kumar Kshitij Patel · Lingxiao Wang · Blake Woodworth · Brian Bullins · Nati Srebro
- abstract@[open-review](#): We study the problem of distributed stochastic non-convex optimization with intermittent communication. We consider the full participation setting where \$M\$ machines work in parallel over \$R\$ communication rounds, as well as the partial participation setting where \$m\$ machines are sampled each round. We propose and analyze a new algorithm that improves existing methods by requiring fewer and lighter variance reduction operations. We also present lower bounds, showing our algorithm is either \textit{optimal} or \textit{almost optimal} in most settings. Numerical experiments demonstrate the superior performance of our algorithm.

## [Few-Shot Non-Parametric Learning with Deep Latent Variable Model](#)

- Zhiying Jiang · Yiqin Dai · Ji Xin · Ming Li · Jimmy Lin
- abstract@[open-review](#): Most real-world problems that machine learning algorithms are expected to solve face the situation with 1) unknown data distribution; 2) little domain-specific knowledge; and 3) datasets with limited annotation. We propose Non-Parametric learning by Compression with Latent Variables (NPC-LV), a learning framework for any dataset with abundant unlabeled data but very few labeled ones. By only training a generative model in an unsupervised way, the framework utilizes the data distribution to build a compressor. Using a compressor-based distance metric derived from Kolmogorov complexity, together with few labeled data, NPC-LV classifies without further training. We show that NPC-LV outperforms supervised methods on all three datasets on image classification in low data regime and even outperform semi-supervised learning methods on CIFAR-10. We demonstrate how and when negative evidence lowerbound (nELBO) can be used as an approximate compressed length for classification. By revealing the

correlation between compression rate and classification accuracy, we illustrate that under NPC-LV, the improvement of generative models can enhance downstream classification accuracy.

## [GULP: a prediction-based metric between representations](#)

- Enric Boix-Adsera · Hannah Lawrence · George Stepaniants · Philippe Rigollet
- abstract@[open-review](#): Comparing the representations learned by different neural networks has recently emerged as a key tool to understand various architectures and ultimately optimize them. In this work, we introduce GULP, a family of distance measures between representations that is explicitly motivated by downstream predictive tasks. By construction, GULP provides uniform control over the difference in prediction performance between two representations, with respect to regularized linear prediction tasks. Moreover, it satisfies several desirable structural properties, such as the triangle inequality and invariance under orthogonal transformations, and thus lends itself to data embedding and visualization. We extensively evaluate GULP relative to other methods, and demonstrate that it correctly differentiates between architecture families, converges over the course of training, and captures generalization performance on downstream linear tasks.

## [On the non-universality of deep learning: quantifying the cost of symmetry](#)

- Emmanuel Abbe · Enric Boix-Adsera
- abstract@[open-review](#): We prove a general computational lower bound for learning with neural networks trained by noisy gradient descent (GD). Our result applies whenever GD training is equivariant (true for many standard architectures), and quantifies the alignment needed between architectures and data in order for GD to learn. As applications, (i) we characterize the functions that fully-connected networks can weak-learn on the binary hypercube and unit sphere, demonstrating that depth-2 is as powerful as any other depth for this task; (ii) we extend the merged-staircase necessity result for learning with latent low-dimensional structure [ABM22] to beyond the mean-field regime. Our techniques extend to stochastic gradient descent (SGD), for which we show nontrivial hardness results for learning with fully-connected networks, based on cryptographic assumptions.

## [Embracing Consistency: A One-Stage Approach for Spatio-Temporal Video Grounding](#)

- Yang Jin · Yongzhi Li · Zehuan Yuan · Yadong Mu
- abstract@[open-review](#): Spatio-Temporal video grounding (STVG) focuses on retrieving the spatio-temporal tube of a specific object depicted by a free-form textual expression. Existing approaches mainly treat this complicated task as a parallel frame-grounding problem and thus suffer from two types of inconsistency drawbacks: feature alignment inconsistency and prediction inconsistency. In this paper, we present an end-to-end one-stage framework, termed Spatio-Temporal Consistency-Aware Transformer (STCAT), to alleviate these issues. Specially, we introduce a novel multi-modal template as the global objective to address this task, which explicitly constricts the grounding region and associates the predictions among all video frames. Moreover, to generate the above template under sufficient video-textual perception, an encoder-decoder architecture is proposed for effective global context modeling. Thanks to these critical designs, STCAT enjoys more consistent cross-modal feature alignment and tube prediction without reliance on any pre-trained object detectors. Extensive experiments show that our method outperforms previous state-of-the-arts with clear margins on two challenging video benchmarks (VidSTG and HC-STVG), illustrating the superiority of the proposed framework to better understanding the association between vision and natural language.

## [Fair and Optimal Decision Trees: A Dynamic Programming Approach](#)

- Jacobus van der Linden · Mathijs de Weerdt · Emir Demirović‡
- abstract@[open-review](#): Interpretable and fair machine learning models are required for many applications, such as credit assessment and in criminal justice. Decision trees offer inherent interpretability, especially when they are small. Optimal decision trees are of particular interest because they offer the best performance possible for a given size. However, state-of-the-art algorithms for fair and optimal decision trees have scalability issues, often requiring several hours to find such trees even for small datasets. In contrast to these state-of-the-art methods that use mixed integer programming, we propose a method that exploits the tree structure using dynamic programming. A key component in our method is a new pruning mechanism that reduces the search space by comparing partial solutions based on upper and lower bounds on their final fairness values. As a result, our model can find fair and optimal trees several orders of magnitude faster than previous methods, also for larger datasets that were previously beyond reach. Moreover, we show that with this substantial improvement our method can find the full Pareto front in the trade-off between accuracy and fairness.

## [Optimal Dynamic Regret in LQR Control](#)

- Dheeraj Baby · Yu-Xiang Wang
- abstract@[open-review](#): We consider the problem of nonstochastic control with a sequence of quadratic losses, i.e., LQR control. We provide an efficient online algorithm that achieves an optimal dynamic (policy) regret of  $\tilde{O}(n^{1/3} \mathcal{TV}(M_{1:n})^{2/3} \vee 1)$ , where  $\mathcal{TV}(M_{1:n})$  is the total variation of any oracle sequence of disturbance action policies parameterized by  $M_1, \dots, M_n$  --- chosen in hindsight to cater to unknown nonstationarity. The rate improves the best known rate of  $\tilde{O}(\sqrt{n} (\mathcal{TV}(M_{1:n}) + 1))$  for general convex losses and is information-theoretically optimal for LQR. Main technical components include the reduction of LQR to online linear regression with delayed feedback due to Foster & Simchowitz 2020, as well as a new proper learning algorithm with an optimal  $\tilde{O}(n^{1/3})$  dynamic regret on a family of "minibatched" quadratic losses, which could be of independent interest.

## [The Implicit Delta Method](#)

- Nathan Kallus · James McInerney
- abstract@[open-review](#): Uncertainty quantification is a crucial part of drawing credible conclusions from predictive models, whether concerned about the prediction at a given point or any downstream evaluation that uses the model as input. When the predictive model is simple and its evaluation differentiable, this task is solved by the delta method, where we propagate the asymptotically-normal uncertainty in the predictive model through the evaluation to compute standard errors and Wald confidence intervals. However, this becomes difficult when the model and/or evaluation becomes more complex. Remedies include the bootstrap, but it can be computationally infeasible when training the model even once is costly. In this paper, we propose an alternative, the implicit delta method, which works by infinitesimally regularizing the training loss of the predictive model to automatically assess downstream uncertainty. We show that the change in the evaluation due to regularization is consistent for the asymptotic variance of the evaluation estimator, even when the infinitesimal change is approximated by a finite difference. This provides both a reliable quantification of uncertainty in terms of standard errors as well as permits the construction of calibrated confidence intervals. We discuss connections to other approaches to uncertainty quantification, both Bayesian and frequentist, and demonstrate our approach empirically.

## [Meta-ticket: Finding optimal subnetworks for few-shot learning within randomly initialized neural networks](#)

- Daiki Chijiwa · Shin'ya Yamaguchi · Atsutoshi Kumagai · Yasutoshi Ida
- abstract@[open-review](#): Few-shot learning for neural networks (NNs) is an important problem that aims to train NNs with a few data. The main challenge is how to avoid overfitting since over-parameterized NNs can easily overfit to such small dataset. Previous work (e.g. MAML by Finn et al. 2017) tackles this challenge by meta-learning, which learns how to learn from a few data by using various tasks. On the other hand, one conventional approach to avoid

overfitting is restricting hypothesis spaces by endowing sparse NN structures like convolution layers in computer vision. However, although such manually-designed sparse structures are sample-efficient for sufficiently large datasets, they are still insufficient for few-shot learning. Then the following questions naturally arise: (1) Can we find sparse structures effective for few-shot learning by meta-learning? (2) What benefits will it bring in terms of meta-generalization? In this work, we propose a novel meta-learning approach, called Meta-ticket, to find optimal sparse subnetworks for few-shot learning within randomly initialized NNs. We empirically validated that Meta-ticket successfully discover sparse subnetworks that can learn specialized features for each given task. Due to this task-wise adaptation ability, Meta-ticket achieves superior meta-generalization compared to MAML-based methods especially with large NNs.

## [Finding and Listing Front-door Adjustment Sets](#)

- Hyunchai Jeong · Jin Tian · Elias Bareinboim
- abstract@[open-review](#): Identifying the effects of new interventions from data is a significant challenge found across a wide range of the empirical sciences. A well-known strategy for identifying such effects is Pearl's front-door (FD) criterion. The definition of the FD criterion is declarative, only allowing one to decide whether a specific set satisfies the criterion. In this paper, we present algorithms for finding and enumerating possible sets satisfying the FD criterion in a given causal diagram. These results are useful in facilitating the practical applications of the FD criterion for causal effects estimation and helping scientists to select estimands with desired properties, e.g., based on cost, feasibility of measurement, or statistical power.

## [Task-Free Continual Learning via Online Discrepancy Distance Learning](#)

- Fei Ye · Adrian G. Bors
- abstract@[open-review](#): Learning from non-stationary data streams, also called Task-Free Continual Learning (TFCL) remains challenging due to the absence of explicit task information. Although there are some recently proposed algorithms for TFCL, these methods lack theoretical guarantees. Moreover, there are no theoretical studies for forgetting analysis of TFCL. This paper develops a new theoretical analysis framework that derives generalization bounds based on the discrepancy distance between the visited samples and the entire information made available for training the model. This analysis provides new insights into the forgetting behaviour in classification tasks. Inspired by this theoretical model, we propose a new approach enabled with the dynamic component expansion mechanism for a mixture model, namely Online Discrepancy Distance Learning (ODDL). ODDL estimates the discrepancy between the current memory and the already accumulated knowledge as the expansion signal to ensure a compact network architecture with optimal performance. We then propose a new sample selection approach that selectively stores the samples into the memory buffer through the discrepancy-based measure, further improving the performance. We perform several TFCL experiments with the proposed methodology, which demonstrate that the proposed approach achieves the state of the art performance.

## [The Neural Testbed: Evaluating Joint Predictions](#)

- Ian Osband · Zheng Wen · Seyed Mohammad Asghari · Vikranth Dwaracherla · Xiuyuan Lu · MORTEZA IBRAHIMI · Dieterich Lawson · Botao Hao · Brendan O'Donoghue · Benjamin Van Roy
- abstract@[open-review](#): Predictive distributions quantify uncertainties ignored by point estimates. This paper introduces The Neural Testbed: an open source benchmark for controlled and principled evaluation of agents that generate such predictions. Crucially, the testbed assesses agents not only on the quality of their marginal predictions per input, but also on their joint predictions across many inputs. We evaluate a range of agents using a simple neural network data generating process. Our results indicate that some popular Bayesian deep learning agents do not fare well with joint predictions, even when they can produce accurate marginal predictions. We also show that the quality of joint predictions drives performance in downstream decision tasks. We find these results are robust across choice a wide range of generative models, and highlight the practical importance of joint predictions to the community.

## [A simple but strong baseline for online continual learning: Repeated Augmented Rehearsal](#)

- Yaqian Zhang · Bernhard Pfahringer · Eibe Frank · Albert Bifet · Nick Jin Sean Lim · Yunzhe Jia
- abstract@[open-review](#): Online continual learning (OCL) aims to train neural networks incrementally from a non-stationary data stream with a single pass through data. Rehearsal-based methods attempt to approximate the observed input distributions over time with a small memory and revisit them later to avoid forgetting. Despite its strong empirical performance, rehearsal methods still suffer from a poor approximation of past data's loss landscape with memory samples. This paper revisits the rehearsal dynamics in online settings. We provide theoretical insights on the inherent memory overfitting risk from the viewpoint of biased and dynamic empirical risk minimization, and examine the merits and limits of repeated rehearsal. Inspired by our analysis, a simple and intuitive baseline, Repeated Augmented Rehearsal (RAR), is designed to address the underfitting-overfitting dilemma of online rehearsals. Surprisingly, across four rather different OCL benchmarks, this simple baseline outperforms vanilla rehearsal by 9%-19% and also significantly improves state-of-the-art rehearsal-based methods MIR, ASER, and SCR. We also demonstrate that RAR successfully achieves an accurate approximation of the loss landscape of past data and high-loss ridge aversion in its learning trajectory. Extensive ablation studies are conducted to study the interplay between repeated and augmented rehearsal and reinforcement learning (RL) is applied to dynamically adjust the hyperparameters of RAR to balance the stability-plasticity trade-off online.

## [Seeing the forest and the tree: Building representations of both individual and collective dynamics with transformers](#)

- Ran Liu · Mehdi Azabou · Max Dabagia · Jingyun Xiao · Eva Dyer
- abstract@[open-review](#): Complex time-varying systems are often studied by abstracting away from individual components and their dynamics and building a model of the population-level dynamics from the start. However, when building a collective description, it can be easy to lose sight of each individual and how different individuals contribute to the larger picture. Here, we present a novel transformer architecture for learning from time-varying data by building descriptions of both the individual as well as the collective population dynamics. Rather than combining many individuals into our model at the onset, we develop a separable architecture that operates on individual time-series first before passing them forward; this induces a permutation-invariance property and can be used to transfer across systems of different size and order. After demonstrating that our model can be applied to successfully recover complex interactions and dynamics in many-body systems, we apply our approach to populations of neurons in the nervous system. On neural activity datasets, we show that our multi-scale transformer not only yields robust decoding performance, but also provide impressive performance in transfer. Our results show that it is possible to learn from neurons in one animal's brain and transfer the model on neurons in a different animal's brain, with interpretable neuron correspondence across sets and animals. This finding opens up a new path to decode from and represent large collections of neurons.

## [Sample-Then-Optimize Batch Neural Thompson Sampling](#)

- Zhongxiang Dai · YAO SHU · Bryan Kian Hsiang Low · Patrick Jaillet
- abstract@[open-review](#): Bayesian optimization (BO), which uses a Gaussian process (GP) as a surrogate to model its objective function, is popular for black-box optimization. However, due to the limitations of GPs, BO underperforms in some problems such as those with categorical, high-dimensional or image inputs. To this end, recent works have used the highly expressive neural networks (NNs) as the surrogate model and derived theoretical guarantees using the theory of neural tangent kernel (NTK). However, these works suffer from the limitations of the requirement to invert an extremely large parameter matrix and the restriction to the sequential (rather than batch) setting. To overcome these limitations, we introduce two algorithms based on the Thompson sampling (TS) policy named Sample-Then-Optimize Batch Neural TS (STO-BNTS) and STO-BNTS-Linear. To choose an input query, we only need to train an NN (resp. a linear model) and then choose the query by maximizing the trained NN (resp. linear model), which is equivalently sampled from the GP posterior with the NTK as the kernel function. As a result, our algorithms sidestep the need to invert the large parameter matrix yet

still preserve the validity of the TS policy. Next, we derive regret upper bounds for our algorithms with batch evaluations, and use insights from batch BO and NTK to show that they are asymptotically no-regret under certain conditions. Finally, we verify their empirical effectiveness using practical AutoML and reinforcement learning experiments.

## [Cross-modal Learning for Image-Guided Point Cloud Shape Completion](#)

- Emanuele Aiello · Diego Valsesia · Enrico Magli
- abstract@[open-review](#): In this paper we explore the recent topic of point cloud completion, guided by an auxiliary image. We show how it is possible to effectively combine the information from the two modalities in a localized latent space, thus avoiding the need for complex point cloud reconstruction methods from single views used by the state-of-the-art. We also investigate a novel self-supervised setting where the auxiliary image provides a supervisory signal to the training process by using a differentiable renderer on the completed point cloud to measure fidelity in the image space. Experiments show significant improvements over state-of-the-art supervised methods for both unimodal and multimodal completion. We also show the effectiveness of the self-supervised approach which outperforms a number of supervised methods and is competitive with the latest supervised models only exploiting point cloud information.

## [Fine-Tuning Pre-Trained Language Models Effectively by Optimizing Subnetworks Adaptively](#)

- Haojie Zhang · Ge Li · Jia Li · Zhongjin Zhang · YUQI ZHU · Zhi Jin
- abstract@[open-review](#): Large-scale pre-trained language models have achieved impressive results on a wide range of downstream tasks recently. However, fine-tuning an extremely large-scale pre-trained language model on limited target datasets is often plagued by overfitting and representation degradation. In this paper, we propose a Dynamic Parameter Selection (DPS) algorithm for the large-scale pre-trained models during fine-tuning, which adaptively selects a more promising subnetwork to perform staged updates based on gradients of back-propagation. Experiments on the GLUE benchmark show that DPS outperforms previous fine-tuning methods in terms of overall performance and stability, and consistently achieves better results with variable pre-trained language models. In addition, DPS brings a large magnitude of improvement in out-of-domain transferring experiments and low-resource scenarios, which shows that it can maintain stable general contextual features and reduce the representation collapse.

## [Nonnegative Tensor Completion via Integer Optimization](#)

- Caleb Bugg · Chen Chen · Anil Aswani
- abstract@[open-review](#): Unlike matrix completion, tensor completion does not have an algorithm that is known to achieve the information-theoretic sample complexity rate. This paper develops a new algorithm for the special case of completion for nonnegative tensors. We prove that our algorithm converges in a linear (in numerical tolerance) number of oracle steps, while achieving the information-theoretic rate. Our approach is to define a new norm for nonnegative tensors using the gauge of a particular 0-1 polytope; integer linear programming can, in turn, be used to solve linear separation problems over this polytope. We combine this insight with a variant of the Frank-Wolfe algorithm to construct our numerical algorithm, and we demonstrate its effectiveness and scalability through computational experiments using a laptop on tensors with up to one-hundred million entries.

## [Equivariant Networks for Crystal Structures](#)

- Oumar Kaba · Siamak Ravanbakhsh
- abstract@[open-review](#): Supervised learning with deep models has tremendous potential for applications in materials science. Recently, graph neural networks have been used in this context, drawing direct inspiration from models for molecules. However, materials are typically much more structured than molecules, which is a feature that these models do not leverage. In this work, we introduce a class of models that are equivariant with respect to crystalline symmetry groups. We do this by defining a generalization of the message passing operations that can be used with more general permutation groups, or that can alternatively be seen as defining an expressive convolution operation on the crystal graph. Empirically, these models achieve competitive results with state-of-the-art on the Materials Project dataset.

## [Simple and Optimal Greedy Online Contention Resolution Schemes](#)

- Vasilis Livanos
- abstract@[open-review](#): Matching based markets, like ad auctions, ride-sharing, and eBay, are inherently online and combinatorial, and therefore have been extensively studied under the lens of online stochastic combinatorial optimization models. The general framework that has emerged uses Contention Resolution Schemes (CRSs) introduced by Chekuri, Vondrák, and Zenklusen for combinatorial problems, where one first obtains a fractional solution to a (continuous) relaxation of the objective, and then proceeds to round it. When the order of rounding is controlled by an adversary, it is called an Online Contention Resolution Scheme (OCRSs), which has been successfully applied in online settings such as posted-price mechanisms, prophet inequalities and stochastic probing. The study of greedy OCRSs against an almighty adversary has emerged as one of the most interesting problems since it gives a simple-to-implement scheme against the worst possible scenario. Intuitively, a greedy OCRS has to make all its decisions before the online process starts. We present simple  $\$1/e\$$  - selectable greedy OCRSs for the single-item setting, partition matroids, and transversal matroids. This improves upon the previous state-of-the-art greedy OCRSs of [FSZ16] that achieves  $\$1/4\$$  for these constraints. Finally, we show that no better competitive ratio than  $\$1/e\$$  is possible, making our greedy OCRSs the best possible.

## [From Gradient Flow on Population Loss to Learning with Stochastic Gradient Descent](#)

- Christopher De Sa · Satyen Kale · Jason Lee · Ayush Sekhari · Karthik Sridharan
- abstract@[open-review](#): Stochastic Gradient Descent (SGD) has been the method of choice for learning large-scale non-convex models. While a general analysis of when SGD works has been elusive, there has been a lot of recent progress in understanding the convergence of Gradient Flow (GF) on the population loss, partly due to the simplicity that a continuous-time analysis buys us. An overarching theme of our paper is providing general conditions under which SGD converges, assuming that GF on the population loss converges. Our main tool to establish this connection is a general \textit{converse Lyapunov} like theorem, which implies the existence of a Lyapunov potential under mild assumptions on the rates of convergence of GF. In fact, using these potentials, we show a one-to-one correspondence between rates of convergence of GF and geometrical properties of the underlying objective. When these potentials further satisfy certain self-bounding properties, we show that they can be used to provide a convergence guarantee for Gradient Descent (GD) and SGD (even when the GF path and GD/SGD paths are quite far apart). It turns out that these self-bounding assumptions are in a sense also necessary for GD/SGD to work. Using our framework, we provide a unified analysis for GD/SGD not only for classical settings like convex losses, or objectives that satisfy PL/ KL properties, but also for more complex problems including Phase Retrieval and Matrix sq-root, and extending the results in the recent work of Chatterjee 2022.

## [LieGG: Studying Learned Lie Group Generators](#)

- Anna Sepliarskaia · Artem Moskalev · Ivan Sosnovik · Arnold Smeulders
- abstract@[open-review](#): Symmetries built into the network have appeared to be very beneficial for a wide range of tasks as it saves data to learn them. We depart from the position that when the symmetries are not built into a model *a priori*, it is advantageous for robust networks to learn the symmetries directly from the data to fit the task function. In this paper, we present a method to extract symmetries learned by the neural network and to evaluate the

degree to which the network is invariant to them. With our method, we are able to explicitly retrieve learned invariances in a form of the generators of corresponding Lie-groups without prior knowledge of the symmetries in the data. We use the proposed method to study how symmetrical properties depend on the network's parameterization and configuration. We found that the ability of the network to learn symmetries generalizes over a wide range of architectures. However, the quality of the learned symmetries depends on the depth and number of parameters.

## [Logical Activation Functions: Logit-space equivalents of Probabilistic Boolean Operators](#)

- Scott Lowe · Robert Earle · Jason d'Eon · Thomas Trappenberg · Sageev Oore
- abstract@[open-review](#): The choice of activation functions and their motivation is a long-standing issue within the neural network community. Neuronal representations within artificial neural networks are commonly understood as logits, representing the log-odds score of presence of features within the stimulus. We derive logit-space operators equivalent to probabilistic Boolean logic-gates AND, OR, and XNOR for independent probabilities. Such theories are important to formalize more complex dendritic operations in real neurons, and these operations can be used as activation functions within a neural network, introducing probabilistic Boolean-logic as the core operation of the neural network. Since these functions involve taking multiple exponents and logarithms, they are computationally expensive and not well suited to be directly used within neural networks. Consequently, we construct efficient approximations named  $\text{AND} \setminus \text{AIL}$  (*the AND operator Approximate for Independent Logits*),  $\text{OR} \setminus \text{AIL}$ , and  $\text{XNOR} \setminus \text{AIL}$ , which utilize only comparison and addition operations, have well-behaved gradients, and can be deployed as activation functions in neural networks. Like MaxOut,  $\text{AND} \setminus \text{AIL}$  and  $\text{OR} \setminus \text{AIL}$  are generalizations of ReLU to two-dimensions. While our primary aim is to formalize dendritic computations within a logit-space probabilistic-Boolean framework, we deploy these new activation functions, both in isolation and in conjunction to demonstrate their effectiveness on a variety of tasks including image classification, transfer learning, abstract reasoning, and compositional zero-shot learning.

## [Neural Stochastic PDEs: Resolution-Invariant Learning of Continuous Spatiotemporal Dynamics](#)

- Cristopher Salvi · Maud Lemercier · Andris Gerasimovics
- abstract@[open-review](#): Stochastic partial differential equations (SPDEs) are the mathematical tool of choice for modelling spatiotemporal PDE-dynamics under the influence of randomness. Based on the notion of mild solution of an SPDE, we introduce a novel neural architecture to learn solution operators of PDEs with (possibly stochastic) forcing from partially observed data. The proposed Neural SPDE model provides an extension to two popular classes of physics-inspired architectures. On the one hand, it extends Neural CDEs and variants -- continuous-time analogues of RNNs -- in that it is capable of processing incoming sequential information arriving at arbitrary spatial resolutions. On the other hand, it extends Neural Operators -- generalizations of neural networks to model mappings between spaces of functions -- in that it can parameterize solution operators of SPDEs depending simultaneously on the initial condition and a realization of the driving noise. By performing operations in the spectral domain, we show how a Neural SPDE can be evaluated in two ways, either by calling an ODE solver (emulating a spectral Galerkin scheme), or by solving a fixed point problem. Experiments on various semilinear SPDEs, including the stochastic Navier-Stokes equations, demonstrate how the Neural SPDE model is capable of learning complex spatiotemporal dynamics in a resolution-invariant way, with better accuracy and lighter training data requirements compared to alternative models, and up to 3 orders of magnitude faster than traditional solvers.

## [Fairness in Federated Learning via Core-Stability](#)

- Bhaskar Ray Chaudhury · Linyi Li · Mintong Kang · Bo Li · Ruta Mehta
- abstract@[open-review](#): Federated learning provides an effective paradigm to jointly optimize a model benefited from rich distributed data while protecting data privacy. Nonetheless, the heterogeneity nature of distributed data, especially in the non-IID setting, makes it challenging to define and ensure fairness among local agents. For instance, it is intuitively ``unfair'' for agents with data of high quality to sacrifice their performance due to other agents with low quality data. Currently popular egalitarian and weighted equity-based fairness measures suffer from the aforementioned pitfall. In this work, we aim to formally represent this problem and address these fairness issues using concepts from co-operative game theory and social choice theory. We model the task of learning a shared predictor in the federated setting as a fair public decision making problem, and then define the notion of core-stable fairness: Given  $N$  agents, there is no subset of agents  $S$  that can benefit significantly by forming a coalition among themselves based on their utilities  $U_N$  and  $U_S$  (i.e.,  $(\sum_{i \in S} U_i) / |S| \geq \sum_{i \in N \setminus S} U_i$ ). Core-stable predictors are robust to low quality local data from some agents, and additionally they satisfy Proportionality (each agent gets at least  $1/n$  fraction of the best utility that she can get from any predictor) and Pareto-optimality (there exists no model that can increase the utility of an agent without decreasing the utility of another), two well sought-after fairness and efficiency notions within social choice. We then propose an efficient federated learning protocol CoreFed to optimize a core stable predictor. CoreFed determines a core-stable predictor when the loss functions of the agents are convex. CoreFed also determines approximate core-stable predictors when the loss functions are not convex, like smooth neural networks. We further show the existence of core-stable predictors in more general settings using Kakutani's fixed point theorem. Finally, we empirically validate our analysis on two real-world datasets, and we show that CoreFed achieves higher core-stability fairness than FedAvg while maintaining similar accuracy.

## [MissDAG: Causal Discovery in the Presence of Missing Data with Continuous Additive Noise Models](#)

- Erdun Gao · Ignavier Ng · Mingming Gong · Li Shen · Wei Huang · Tongliang Liu · Kun Zhang · Howard Bondell
- abstract@[open-review](#): State-of-the-art causal discovery methods usually assume that the observational data is complete. However, the missing data problem is pervasive in many practical scenarios such as clinical trials, economics, and biology. One straightforward way to address the missing data problem is first to impute the data using off-the-shelf imputation methods and then apply existing causal discovery methods. However, such a two-step method may suffer from suboptimality, as the imputation algorithm may introduce bias for modeling the underlying data distribution. In this paper, we develop a general method, which we call MissDAG, to perform causal discovery from data with incomplete observations. Focusing mainly on the assumptions of ignorable missingness and the identifiable additive noise models (ANMs), MissDAG maximizes the expected likelihood of the visible part of observations under the expectation-maximization (EM) framework. In the E-step, in cases where computing the posterior distributions of parameters in closed-form is not feasible, Monte Carlo EM is leveraged to approximate the likelihood. In the M-step, MissDAG leverages the density transformation to model the noise distributions with simpler and specific formulations by virtue of the ANMs and uses a likelihood-based causal discovery algorithm with directed acyclic graph constraint. We demonstrate the flexibility of MissDAG for incorporating various causal discovery algorithms and its efficacy through extensive simulations and real data experiments.

## [Near-Optimal Regret for Adversarial MDP with Delayed Bandit Feedback](#)

- Tiancheng Jin · Tal Lancewicki · Haipeng Luo · Yishay Mansour · Aviv Rosenberg
- abstract@[open-review](#): The standard assumption in reinforcement learning (RL) is that agents observe feedback for their actions immediately. However, in practice feedback is often observed in delay. This paper studies online learning in episodic Markov decision process (MDP) with unknown transitions, adversarially changing costs, and unrestricted delayed bandit feedback. More precisely, the feedback for the agent in episode  $k$  is revealed only in the end of episode  $k + d^k$ , where the delay  $d^k$  can be changing over episodes and chosen by an oblivious adversary. We present the first algorithms that achieve near-optimal  $\sqrt{K + D}$  regret, where  $K$  is the number of episodes and  $D = \sum_{k=1}^K d^k$  is the total delay, significantly improving upon the best known regret bound of  $(K + D)^{2/3}$ .

## [AMP: Automatically Finding Model Parallel Strategies with Heterogeneity Awareness](#)

- Dacheng Li · Hongyi Wang · Eric Xing · Hao Zhang
- abstract@[open-review](#): Scaling up model sizes can lead to fundamentally new capabilities in many machine learning (ML) tasks. However, training big models requires strong distributed system expertise to carefully design model-parallel execution strategies that suit with the model architectures and cluster setups. In this paper, we develop AMP, a framework to automatically derive such strategies. AMP identifies a valid space of model parallelism strategies, and efficiently searches the space for high-performed strategies, by leveraging a cost model designed to capture the heterogeneity of the model and cluster specifications. Unlike existing methods, AMP is specifically tailored to support complex models composed of uneven layers and cluster setups with more heterogeneous accelerators and bandwidth. We evaluate AMP on popular models and cluster setups from public clouds and show that: AMP returns parallel strategies that match the expert-tuned strategies on typical cluster setups. On heterogeneous clusters or models with heterogeneous architectures, AMP demonstrates up to 1.37x and 1.76x higher throughput than state-of-the-art model-parallel systems, respectively.

## [New Definitions and Evaluations for Saliency Methods: Staying Intrinsic, Complete and Sound](#)

- Arushi Gupta · Nikunj Saunshi · Dingli Yu · Kaifeng Lyu · Sanjeev Arora
- abstract@[open-review](#): Saliency methods compute heat maps that highlight portions of an input that were most \em important} for the label assigned to it by a deep net. Evaluations of saliency methods convert this heat map into a new \em masked input} by retaining the \$k\$ highest-ranked pixels of the original input and replacing the rest with \textquotedblleft uninformative\textquotedblright pixels, and checking if the net's output is mostly unchanged. This is usually seen as an \em explanation} of the output, but the current paper highlights reasons why this inference of causality may be suspect. Inspired by logic concepts of \em completeness \& soundness}, it observes that the above type of evaluation focuses on completeness of the explanation, but ignores soundness. New evaluation metrics are introduced to capture both notions, while staying in an \em intrinsic} framework---i.e., using the dataset and the net, but no separately trained nets, human evaluations, etc. A simple saliency method is described that matches or outperforms prior methods in the evaluations. Experiments also suggest new intrinsic justifications, based on soundness, for popular heuristic tricks such as TV regularization and upsampling.

## [Hand-Object Interaction Image Generation](#)

- Hezhen Hu · Weilun Wang · Wengang Zhou · Houqiang Li
- abstract@[open-review](#): In this work, we are dedicated to a new task, i.e., hand-object interaction image generation, which aims to conditionally generate the hand-object image under the given hand, object and their interaction status. This task is challenging and research-worthy in many potential application scenarios, such as AR/VR games and online shopping, etc. To address this problem, we propose a novel HOGAN framework, which utilizes the expressive model-aware hand-object representation and leverages its inherent topology to build the unified surface space. In this space, we explicitly consider the complex self- and mutual occlusion during interaction. During final image synthesis, we consider different characteristics of hand and object and generate the target image in a split-and-combine manner. For evaluation, we build a comprehensive protocol to access both the fidelity and structure preservation of the generated image. Extensive experiments on two large-scale datasets, i.e., HO3Dv3 and DexYCB, demonstrate the effectiveness and superiority of our framework both quantitatively and qualitatively.

## [Structuring Representations using Geometric Invariants](#)

- Mehran Shakerinava · Arnab Mondal · Siamak Ravanbakhsh
- abstract@[open-review](#): Different geometries can be defined by their invariants; for example, distances and angles are preserved in Euclidean and Conformal geometries, respectively. We propose to structure equivariant representations using loss functions based on these invariants. This approach requires no prior knowledge about the correspondence of potentially complex transformations in the input space and the symmetry group. We provide a general recipe based on polynomial invariants of groups, and consider examples within Lie groups such as the Euclidean and Orthogonal groups, as well as finite groups, such as the symmetric group. We also show the feasibility of learning disentangled representations using this approach and provide favorable qualitative and quantitative results on downstream tasks, including world modeling and reinforcement learning.

## [Learning to Re-weight Examples with Optimal Transport for Imbalanced Classification](#)

- Dandan Guo · Zhuo Li · meixi zheng · He Zhao · Mingyuan Zhou · Hongyuan Zha
- abstract@[open-review](#): Imbalanced data pose challenges for deep learning based classification models. One of the most widely-used approaches for tackling imbalanced data is re-weighting, where training samples are associated with different weights in the loss function. Most of existing re-weighting approaches treat the example weights as the learnable parameter and optimize the weights on the meta set, entailing expensive bilevel optimization. In this paper, we propose a novel re-weighting method based on optimal transport (OT) from a distributional point of view. Specifically, we view the training set as an imbalanced distribution over its samples, which is transported by OT to a balanced distribution obtained from the meta set. The weights of the training samples are the probability mass of the imbalanced distribution and learned by minimizing the OT distance between the two distributions. Compared with existing methods, our proposed one disengages the dependence of the weight learning on the concerned classifier at each iteration. Experiments on image, text and point cloud datasets demonstrate that our proposed re-weighting method has excellent performance, achieving state-of-the-art results in many cases and providing a promising tool for addressing the imbalanced classification issue.

## [Parameter-free Regret in High Probability with Heavy Tails](#)

- Juijia Zhang · Ashok Cutkosky
- abstract@[open-review](#): We present new algorithms for online convex optimization over unbounded domains that obtain parameter-free regret in high-probability given access only to potentially heavy-tailed subgradient estimates. Previous work in unbounded domains considers only in-expectation results for sub-exponential subgradients. Unlike in the bounded domain case, we cannot rely on straight-forward martingale concentration due to potentially exponentially large iterates produced by the algorithm. We develop new techniques based on novel regularizers to overcome these problems.

## [Root Cause Analysis of Failures in Microservices through Causal Discovery](#)

- Azam Ikram · Sarthak Chakraborty · Subrata Mitra · Shiv Saini · Saurabh Bagchi · Murat Kocaoglu
- abstract@[open-review](#): The majority of the production web services use a large number of smaller sub-components (called microservices) that interact with each other, in the form of a complex graph to provide the overall functionality to the user. While modularity of the microservice architecture is beneficial for rapid software development and deployment, maintaining and debugging such a system quickly in case of a failure or a poor performance is challenging. We propose a scalable algorithm for quickly detecting the root cause of failure in complex microservice architectures. The key ideas behind our novel localized hierarchical learning approach are: (1) to treat a failure as an intervention on the root cause to learn the underlying causal structure, (2) only learn the portion of the causal graph related to the root cause, thus avoiding a large number of costly conditional independence tests, and (3) hierarchically learn the relevant causal graph. The proposed technique is highly scalable and can produce useful insights about the root cause, while the use of traditional techniques becomes infeasible due to high computation time. Our solution is application agnostic and only relies on the data collected for diagnosis. For our evaluation, we compare our proposed solution with a modified version of the PC algorithm and the state-of-the-art for root cause analysis. The results show a significant improvement in top-\$k\$ precision while keeping the execution time reasonable.

## [Streaming Radiance Fields for 3D Video Synthesis](#)

- Lingzhi LI · Zhen Shen · zhongshu wang · Li Shen · Ping Tan
- abstract@[open-review](#): We present an explicit-grid based method for efficiently reconstructing streaming radiance fields for novel view synthesis of real world dynamic scenes. Instead of training a single model that combines all the frames, we formulate the dynamic modeling problem with an incremental learning paradigm in which per-frame model difference is trained to complement the adaption of a base model on the current frame. By exploiting the simple yet effective strategy of tuning with narrow bands, the proposed method achieves a feasible framework for handling on-the-fly video sequences with high training efficiency. The storage overhead induced by using explicit grid can be significantly reduced by using model difference based compression. We also introduce a curriculum learning strategy to further accelerate model optimization for each frame. Experiments on challenging video sequences demonstrate that our approach is capable of achieving a training speed of 10 seconds per-frame with competitive rendering quality, which attains \$1500 \times\$ speedup over the state-of-the-art implicit methods.

## [A Unified Convergence Theorem for Stochastic Optimization Methods](#)

- Xiao Li · Andre Milzarek
- abstract@[open-review](#): In this work, we provide a fundamental unified convergence theorem used for deriving expected and almost sure convergence results for a series of stochastic optimization methods. Our unified theorem only requires to verify several representative conditions and is not tailored to any specific algorithm. As a direct application, we recover expected and almost sure convergence results of the stochastic gradient method (SGD) and random reshuffling (RR) under more general settings. Moreover, we establish new expected and almost sure convergence results for the stochastic proximal gradient method (prox-SGD) and stochastic model-based methods for nonsmooth nonconvex optimization problems. These applications reveal that our unified theorem provides a plugin-type convergence analysis and strong convergence guarantees for a wide class of stochastic optimization methods.

## [Neural Set Function Extensions: Learning with Discrete Functions in High Dimensions](#)

- Nikolaos Karalias · Joshua Robinson · Andreas Loukas · Stefanie Jegelka
- abstract@[open-review](#): Integrating functions on discrete domains into neural networks is key to developing their capability to reason about discrete objects. But, discrete domains are (1) not naturally amenable to gradient-based optimization, and (2) incompatible with deep learning architectures that rely on representations in high-dimensional vector spaces. In this work, we address both difficulties for set functions, which capture many important discrete problems. First, we develop a framework for extending set functions onto low-dimensional continuous domains, where many extensions are naturally defined. Our framework subsumes many well-known extensions as special cases. Second, to avoid undesirable low-dimensional neural network bottlenecks, we convert low-dimensional extensions into representations in high-dimensional spaces, taking inspiration from the success of semidefinite programs for combinatorial optimization. Empirically, we observe benefits of our extensions for unsupervised neural combinatorial optimization, in particular with high-dimensional representations.

## [Counterfactual Temporal Point Processes](#)

- Kimia Noorbakhsh · Manuel Rodriguez
- abstract@[open-review](#): Machine learning models based on temporal point processes are the state of the art in a wide variety of applications involving discrete events in continuous time. However, these models lack the ability to answer counterfactual questions, which are increasingly relevant as these models are being used to inform targeted interventions. In this work, our goal is to fill this gap. To this end, we first develop a causal model of thinning for temporal point processes that builds upon the Gumbel-Max structural causal model. This model satisfies a desirable counterfactual monotonicity condition, which is sufficient to identify counterfactual dynamics in the process of thinning. Then, given an observed realization of a temporal point process with a given intensity function, we develop a sampling algorithm that uses the above causal model of thinning and the superposition theorem to simulate counterfactual realizations of the temporal point process under a given alternative intensity function. Simulation experiments using synthetic and real epidemiological data show that the counterfactual realizations provided by our algorithm may give valuable insights to enhance targeted interventions.

## [Quality Not Quantity: On the Interaction between Dataset Design and Robustness of CLIP](#)

- Thao Nguyen · Gabriel Ilharco · Mitchell Wortsman · Sewoong Oh · Ludwig Schmidt
- abstract@[open-review](#): Web-crawled datasets have enabled remarkable generalization capabilities in recent image-text models such as CLIP (Contrastive Language-Image pre-training) or Flamingo, but little is known about the dataset creation processes. In this work, we introduce a testbed of six publicly available data sources---YFCC, LAION, Conceptual Captions, WIT, RedCaps, Shutterstock---to investigate how pre-training distributions induce robustness in CLIP. We find that the performance of the pre-training data varies substantially across distribution shifts, with no single data source dominating. Moreover, we systematically study the interactions between these data sources and find that mixing multiple sources does not necessarily yield better models, but rather dilutes the robustness of the best individual data source. We complement our empirical findings with theoretical insights from a simple setting, where combining the training data also results in diluted robustness. In addition, our theoretical model provides a candidate explanation for the success of the CLIP-based data filtering technique recently employed in the LAION dataset. Overall our results demonstrate that simply gathering a large amount of data from the web is not the most effective way to build a pre-training dataset for robust generalization, necessitating further study into dataset design.

## [Promising or Elusive? Unsupervised Object Segmentation from Real-world Single Images](#)

- Yafei YANG · Bo Yang
- abstract@[open-review](#): In this paper, we study the problem of unsupervised object segmentation from single images. We do not introduce a new algorithm, but systematically investigate the effectiveness of existing unsupervised models on challenging real-world images. We firstly introduce four complexity factors to quantitatively measure the distributions of object- and scene-level biases in appearance and geometry for datasets with human annotations. With the aid of these factors, we empirically find that, not surprisingly, existing unsupervised models catastrophically fail to segment generic objects in real-world images, although they can easily achieve excellent performance on numerous simple synthetic datasets, due to the vast gap in objectness biases between synthetic and real images. By conducting extensive experiments on multiple groups of ablated real-world datasets, we ultimately find that the key factors underlying the colossal failure of existing unsupervised models on real-world images is the challenging distributions of object- and scene-level biases in appearance and geometry. Because of this, the inductive biases introduced in existing unsupervised models can hardly capture the diverse object distributions. Our research results suggest that future work should exploit more explicit objectness biases in the network design.

## [Amortized Proximal Optimization](#)

- Juhan Bae · Paul Vicol · Jeff Z. HaoChen · Roger Grosse
- abstract@[open-review](#): We propose a framework for online meta-optimization of parameters that govern optimization, called Amortized Proximal Optimization (APO). We first interpret various existing neural network optimizers as approximate stochastic proximal point methods which trade off the current-batch loss with proximity terms in both function space and weight space. The idea behind APO is to amortize the minimization of the proximal point objective by meta-learning the parameters of an update rule. We show how APO can be used to adapt a learning rate or a structured preconditioning matrix. Under appropriate assumptions, APO can recover existing optimizers such as natural gradient descent and KFAC. It enjoys low computational overhead and avoids expensive and numerically sensitive operations required by some second-order optimizers, such as matrix inverses. We empirically test APO for online adaptation of learning rates and structured preconditioning matrices for regression, image reconstruction, image classification, and

natural language translation tasks. Empirically, the learning rate schedules found by APO generally outperform optimal fixed learning rates and are competitive with manually tuned decay schedules. Using APO to adapt a structured preconditioning matrix generally results in optimization performance competitive with second-order methods.

## [Near-Optimal Correlation Clustering with Privacy](#)

- Vincent Cohen-Addad · Chenglin Fan · Silvio Lattanzi · Slobodan Mitrovic · Ashkan Norouzi-Fard · Nikos Parotsidis · Jakub Tarnawski
- abstract@[open-review](#): Correlation clustering is a central problem in unsupervised learning, with applications spanning community detection, duplicate detection, automated labeling and many more. In the correlation clustering problem one receives as input a set of nodes and for each node a list of co-clustering preferences, and the goal is to output a clustering that minimizes the disagreement with the specified nodes' preferences. In this paper, we introduce a simple and computationally efficient algorithm for the correlation clustering problem with provable privacy guarantees. Our additive error is stronger than those obtained in prior work and is optimal up to polylogarithmic factors for fixed privacy parameters.

## [Conditional Meta-Learning of Linear Representations](#)

- Giulia Denevi · Massimiliano Pontil · Carlo Ciliberto
- abstract@[open-review](#): Standard meta-learning for representation learning aims to find a common representation to be shared across multiple tasks. The effectiveness of these methods is often limited when the nuances of the tasks' distribution cannot be captured by a single representation. In this work we overcome this issue by inferring a conditioning function, mapping the tasks' side information (such as the tasks' training dataset itself) into a representation tailored to the task at hand. We study environments in which our conditional strategy outperforms standard meta-learning, such as those in which tasks can be organized in separate clusters according to the representation they share. We then propose a meta-algorithm capable of leveraging this advantage in practice. In the unconditional setting, our method yields a new estimator enjoying faster learning rates and requiring less hyper-parameters to tune than current state-of-the-art methods. Our results are supported by preliminary experiments.

## [Latent-Variable Advantage-Weighted Policy Optimization for Offline Reinforcement Learning](#)

- Xi Chen · Ali Ghadirzadeh · Tianhe Yu · Alex Yuan Gao · Jianhao Wang · Wenzhe Li · Liang Bin · Chelsea Finn · Chongjie Zhang
- abstract@[open-review](#): Offline reinforcement learning methods hold the promise of learning policies from pre-collected datasets without the need to query the environment for new samples. This setting is particularly well-suited for continuous control robotic applications for which online data collection based on trial-and-error is costly and potentially unsafe. In practice, offline datasets are often heterogeneous, i.e., collected in a variety of scenarios, such as data from several human demonstrators or from policies that act with different purposes. Unfortunately, such datasets often contain action distributions with multiple modes and, in some cases, lack a sufficient number of high-reward trajectories, which render offline policy training inefficient. To address this challenge, we propose to leverage latent-variable generative model to represent high-advantage state-action pairs leading to better adherence to data distributions that contributes to solving the task, while maximizing reward via a policy over the latent variable. As we empirically show on a range of simulated locomotion, navigation, and manipulation tasks, our method referred to as latent-variable advantage-weighted policy optimization (LAPO), improves the average performance of the next best-performing offline reinforcement learning methods by 49% on heterogeneous datasets, and by 8% on datasets with narrow and biased distributions.

## [Luckiness in Multiscale Online Learning](#)

- Wouter Koolen · Muriel Părež
- abstract@[open-review](#): Algorithms for full-information online learning are classically tuned to minimize the worst-case regret. Modern algorithms in addition provide tighter guarantees outside the maximally adversarial regime, most notably in the form of constant (pseudo)-regret bounds under statistical margin assumptions. We investigate the multiscale extension of the setting, where the loss ranges of the various experts are vastly different, and the regret w.r.t. each expert needs to scale with its range, instead of the maximum overall range. We develop new algorithms, tuning schemes and analysis techniques, and show that indeed one can combine worst-case robustness with adaptation to easy data at negligible cost. We develop an extension with optimism, and apply it to solving multiscale zero-sum games. We demonstrate in experiments the superior performance of our scale-adaptive algorithm. We discuss the subtle relationship of our results to Freund's 2016 problem.

## [Diverse Weight Averaging for Out-of-Distribution Generalization](#)

- Alexandre Rame · Matthieu Kirchmeyer · Thibaud Rahier · Alain Rakotomamonjy · Patrick Gallinari · Matthieu Cord
- abstract@[open-review](#): Standard neural networks struggle to generalize under distribution shifts in computer vision. Fortunately, combining multiple networks can consistently improve out-of-distribution generalization. In particular, weight averaging (WA) strategies were shown to perform best on the competitive DomainBed benchmark; they directly average the weights of multiple networks despite their nonlinearities. In this paper, we propose Diverse Weight Averaging (DiWA), a new WA strategy whose main motivation is to increase the functional diversity across averaged models. To this end, DiWA averages weights obtained from several independent training runs: indeed, models obtained from different runs are more diverse than those collected along a single run thanks to differences in hyperparameters and training procedures. We motivate the need for diversity by a new bias-variance-covariance-locality decomposition of the expected error, exploiting similarities between WA and standard functional ensembling. Moreover, this decomposition highlights that WA succeeds when the variance term dominates, which we show occurs when the marginal distribution changes at test time. Experimentally, DiWA consistently improves the state of the art on DomainBed without inference overhead.

## [Random Rank: The One and Only Strategyproof and Proportionally Fair Randomized Facility Location Mechanism](#)

- Haris Aziz · Alexander Lam · Mashbat Suzuki · Toby Walsh
- abstract@[open-review](#): Proportionality is an attractive fairness concept that has been applied to a range of problems including the facility location problem, a classic problem in social choice. In our work, we propose a concept called Strong Proportionality, which ensures that when there are two groups of agents at different locations, both groups incur the same total cost. We show that although Strong Proportionality is a well-motivated and basic axiom, there is no deterministic strategyproof mechanism satisfying the property. We then identify a randomized mechanism called Random Rank (which uniformly selects a number  $k$  between  $1$  to  $n$  and locates the facility at the  $k$ 'th highest agent location) which satisfies Strong Proportionality in expectation. Our main theorem characterizes Random Rank as the unique mechanism that achieves universal truthfulness, universal anonymity, and Strong Proportionality in expectation among all randomized mechanisms. Finally, we show via the AverageOrRandomRank mechanism that even stronger ex-post fairness guarantees can be achieved by weakening universal truthfulness to strategyproofness in expectation.

## [Semi-infinitely Constrained Markov Decision Processes](#)

- Liangyu Zhang · Yang Peng · Wenhao Yang · Zhihua Zhang
- abstract@[open-review](#): We propose a generalization of constrained Markov decision processes (CMDPs) that we call the semi-infinitely constrained Markov decision process (SICMDP). Particularly, in a SICMDP model, we impose a continuum of constraints instead of a finite number of constraints as in the case of ordinary CMDPs. We also devise a reinforcement learning algorithm for SICMDPs that we call SI-CRL. We first transform the reinforcement learning problem into a linear semi-infinitely programming (LSIP) problem and then use the dual exchange method in the LSIP literature to

solve it. To the best of our knowledge, we are the first to apply tools from semi-infinitely programming (SIP) to solve reinforcement learning problems. We present theoretical analysis for SI-CRL, identifying its sample complexity and iteration complexity. We also conduct extensive numerical examples to illustrate the SICMDP model and validate the SI-CRL algorithm.

## [A Differentiable Semantic Metric Approximation in Probabilistic Embedding for Cross-Modal Retrieval](#)

- Hao Li · Jingkuan Song · Lianli Gao · Pengpeng Zeng · Haonan Zhang · Gongfu Li
- abstract@[open-review](#): Cross-modal retrieval aims to build correspondence between multiple modalities by learning a common representation space. Typically, an image can match multiple texts semantically, and vice versa, which greatly increases the difficulty of this task. To tackle this problem, probabilistic embeddings are proposed to quantify these many-to-many relationships. However, existing datasets (e.g., MS-COCO) and metrics (e.g., Recall@K) are hard to fully represent these diversity correspondences due to non-exhaustive annotations. Based on this observation, we utilize semantic correlation computed by CIDEr to find the potential correspondence. Then we present an effective metric, named Average Semantic Precision (ASP), which can measure the ranking precision of semantic correlation for retrieval sets. Additionally, we introduce a novel and concise objective, coined Differentiable ASP Approximation (DAA). Concretely, DAA can optimize ASP directly by making the ranking function of ASP differentiable through a sigmoid function. To verify the effectiveness of our approach, extensive experiments are conducted on MS-COCO and CUB Captions, which are commonly used in probabilistic embedding for cross-modal retrieval. The results show that our approach obtains superior performance over the state-of-the-art approaches on all metrics. The code and trained models are released at \url{https://anonymous.4open.science/r/2022-NeurIPS-DAA-4F1F}.

## [Debiased, Longitudinal and Coordinated Drug Recommendation through Multi-Visit Clinic Records](#)

- Hongda Sun · Shufang Xie · Shuqi Li · Yuhua Chen · Ji-Rong Wen · Rui Yan
- abstract@[open-review](#): AI-empowered drug recommendation has become an important task in healthcare research areas, which offers an additional perspective to assist human doctors with more accurate and more efficient drug prescriptions. Generally, drug recommendation is based on patients' diagnosis results in the electronic health records. We assume that there are three key factors to be addressed in drug recommendation: (1) elimination of recommendation bias due to limitations of observable information, (2) better utilization of historical health condition and (3) coordination of multiple drugs to control safety. To this end, we propose DrugRec, a causal inference based drug recommendation model. The causal graphical model can identify and deconfound the recommendation bias with front-door adjustment. Meanwhile, we model the multi-visit in the causal graph to characterize a patient's historical health conditions. Finally, we model the drug-drug interactions (DDIs) as the propositional satisfiability (SAT) problem, and solving the SAT problem can help better coordinate the recommendation. Comprehensive experiment results show that our proposed model achieves state-of-the-art performance on the widely used datasets MIMIC-III and MIMIC-IV, demonstrating the effectiveness and safety of our method.

## [Semantic Exploration from Language Abstractions and Pretrained Representations](#)

- Allison Tam · Neil Rabinowitz · Andrew Lampinen · Nicholas Roy · Stephanie Chan · DJ Strouse · Jane Wang · Andrea Banino · Felix Hill
- abstract@[open-review](#): Effective exploration is a challenge in reinforcement learning (RL). Novelty-based exploration methods can suffer in high-dimensional state spaces, such as continuous partially-observable 3D environments. We address this challenge by defining novelty using semantically meaningful state abstractions, which can be found in learned representations shaped by natural language. In particular, we evaluate vision-language representations, pretrained on natural image captioning datasets. We show that these pretrained representations drive meaningful, task-relevant exploration and improve performance on 3D simulated environments. We also characterize why and how language provides useful abstractions for exploration by considering the impacts of using representations from a pretrained model, a language oracle, and several ablations. We demonstrate the benefits of our approach with on- and off-policy RL algorithms and in two very different task domains---one that stresses the identification and manipulation of everyday objects, and one that requires navigational exploration in an expansive world. Our results suggest that using language-shaped representations could improve exploration for various algorithms and agents in challenging environments.

## [Coresets for Wasserstein Distributionally Robust Optimization Problems](#)

- Ruomin Huang · Jiawei Huang · Wenjie Liu · Hu Ding
- abstract@[open-review](#): Wasserstein distributionally robust optimization (\textsf{WDRO}) is a popular model to enhance the robustness of machine learning with ambiguous data. However, the complexity of \textsf{WDRO} can be prohibitive in practice since solving its minimax'' formulation requires a great amount of computation. Recently, several fast \textsf{WDRO} training algorithms for some specific machine learning tasks (e.g., logistic regression) have been developed. However, the research on designing efficient algorithms for general large-scale \textsf{WDRO}s is still quite limited, to the best of our knowledge. \textit{Coreset} is an important tool for compressing large dataset, and thus it has been widely applied to reduce the computational complexities for many optimization problems. In this paper, we introduce a unified framework to construct the \$\epsilon\$-coreset for the general \textsf{WDRO} problems. Though it is challenging to obtain a conventional coreset for \textsf{WDRO} due to the uncertainty issue of ambiguous data, we show that we can compute adual coreset" by using the strong duality property of \textsf{WDRO}. Also, the error introduced by the dual coreset can be theoretically guaranteed for the original \textsf{WDRO} objective. To construct the dual coreset, we propose a novel grid sampling approach that is particularly suitable for the dual formulation of \textsf{WDRO}. Finally, we implement our coreset approach and illustrate its effectiveness for several \textsf{WDRO} problems in the experiments.

## [Tree Mover's Distance: Bridging Graph Metrics and Stability of Graph Neural Networks](#)

- Ching-Yao Chuang · Stefanie Jegelka
- abstract@[open-review](#): Understanding generalization and robustness of machine learning models fundamentally relies on assuming an appropriate metric on the data space. Identifying such a metric is particularly challenging for non-Euclidean data such as graphs. Here, we propose a pseudometric for attributed graphs, the Tree Mover's Distance (TMD), and study its relation to generalization. Via a hierarchical optimal transport problem, TMD reflects the local distribution of node attributes as well as the distribution of local computation trees, which are known to be decisive for the learning behavior of graph neural networks (GNNs). First, we show that TMD captures properties relevant for graph classification: a simple TMD-SVM can perform competitively with standard GNNs. Second, we relate TMD to generalization of GNNs under distribution shifts, and show that it correlates well with performance drop under such shifts.

## [The Privacy Onion Effect: Memorization is Relative](#)

- Nicholas Carlini · Matthew Jagielski · Chiyuan Zhang · Nicolas Papernot · Andreas Terzis · Florian Tramer
- abstract@[open-review](#): Machine learning models trained on private datasets have been shown to leak their private data. Recent work has found that the average data point is rarely leaked---it is often the outlier samples that are subject to memorization and, consequently, leakage. We demonstrate and analyze an Onion Effect of memorization: removing the "layer" of outlier points that are most vulnerable to a privacy attack exposes a new layer of previously-safe points to the same attack. We perform several experiments that are consistent with this hypothesis. For example, we show that for membership inference attacks, when the layer of easiest-to-attack examples is removed, another layer below becomes easy-to-attack. The existence of this effect has various consequences. For example, it suggests that proposals to defend against memorization without training with rigorous privacy guarantees are unlikely to be effective. Further, it suggests that privacy-enhancing technologies such as machine unlearning could actually harm the privacy of other users.

## Hyperbolic Embedding Inference for Structured Multi-Label Prediction

- Bo Xiong · Michael Cochez · Mojtaba Nayyeri · Steffen Staab
- abstract@[open-review](#): We consider a structured multi-label prediction problem where the labels are organized under implication and mutual exclusion constraints. A major concern is to produce predictions that are logically consistent with these constraints. To do so, we formulate this problem as an embedding inference problem where the constraints are imposed onto the embeddings of labels by geometric construction. Particularly, we consider a hyperbolic Poincaré ball model in which we encode labels as Poincaré hyperplanes that work as linear decision boundaries. The hyperplanes are interpreted as convex regions such that the logical relationships (implication and exclusion) are geometrically encoded using the insideness and disjointedness of these regions, respectively. We show theoretical groundings of the method for preserving logical relationships in the embedding space. Extensive experiments on 12 datasets show 1) significant improvements in mean average precision; 2) lower number of constraint violations; 3) an order of magnitude fewer dimensions than baselines.

## Pre-Trained Image Encoder for Generalizable Visual Reinforcement Learning

- Zhecheng Yuan · Zhengrong Xue · Bo Yuan · Xueqian Wang · YI WU · Yang Gao · Huazhe Xu
- abstract@[open-review](#): Learning generalizable policies that can adapt to unseen environments remains challenging in visual Reinforcement Learning (RL). Existing approaches try to acquire a robust representation via diversifying the appearances of in-domain observations for better generalization. Limited by the specific observations of the environment, these methods ignore the possibility of exploring diverse real-world image datasets. In this paper, we investigate how a visual RL agent would benefit from the off-the-shelf visual representations. Surprisingly, we find that the early layers in an ImageNet pre-trained ResNet model could provide rather generalizable representations for visual RL. Hence, we propose Pre-trained Image Encoder for Generalizable visual reinforcement learning (PIE-G), a simple yet effective framework that can generalize to the unseen visual scenarios in a zero-shot manner. Extensive experiments are conducted on DMControl Generalization Benchmark, DMControl Manipulation Tasks, and Drawer World to verify the effectiveness of PIE-G. Empirical evidence suggests PIE-G improves sample efficiency and significantly outperforms previous state-of-the-art methods in terms of generalization performance. In particular, PIE-G boasts a 55% generalization performance gain on average in the challenging video background setting. Project Page: <https://sites.google.com/view/pie-g/home>.

## DNA: Proximal Policy Optimization with a Dual Network Architecture

- Matthew Aitchison · Penny Sweetser
- abstract@[open-review](#): This paper explores the problem of simultaneously learning a value function and policy in deep actor-critic reinforcement learning models. We find that the common practice of learning these functions jointly is sub-optimal, due to an order-of-magnitude difference in noise levels between these two tasks. Instead, we show that learning these tasks independently, but with a constrained distillation phase, significantly improves performance. Furthermore, we find that the policy learning noise levels can be decreased by using a lower  $\text{variance}$  return estimate. Whereas, the value learning noise level is instead decreased with a lower  $\text{bias}$  estimate. Together these insights inform an extension to Proximal Policy Optimization we call `Dual Network Architecture` (DNA), which significantly outperforms its predecessor. DNA also exceeds the performance of the popular Rainbow DQN algorithm on four of the five environments tested, even under more difficult stochastic control settings.

## Block-Recurrent Transformers

- DeLesley Hutchins · Imanol Schlag · Ethan Dyer · Behnam Neyshabur · Yuhuai Wu
- abstract@[open-review](#): We introduce the Block-Recurrent Transformer, which applies a transformer layer in a recurrent fashion along a sequence, and has linear complexity with respect to sequence length. Our recurrent cell operates on blocks of tokens rather than single tokens during training, and leverages parallel computation within a block in order to make efficient use of accelerator hardware. The cell itself is strikingly simple. It is merely a transformer layer: it uses self-attention and cross-attention to efficiently compute a recurrent function over a large set of state vectors and tokens. Our design was inspired in part by LSTM cells, and it uses LSTM-style gates, but it scales the typical LSTM cell up by several orders of magnitude. Our implementation of recurrence has the same cost in both computation time and parameter count as a conventional transformer layer, but offers dramatically improved perplexity in language modeling tasks over very long sequences. Our model out-performs a long-range Transformer XL baseline by a wide margin, while running twice as fast. We demonstrate its effectiveness on PG19 (books), arXiv papers, and GitHub source code. Our code has been released as open source.

## Denoising Diffusion Restoration Models

- Bahjat Kawar · Michael Elad · Stefano Ermon · Jiaming Song
- abstract@[open-review](#): Many interesting tasks in image restoration can be cast as linear inverse problems. A recent family of approaches for solving these problems uses stochastic algorithms that sample from the posterior distribution of natural images given the measurements. However, efficient solutions often require problem-specific supervised training to model the posterior, whereas unsupervised methods that are not problem-specific typically rely on inefficient iterative methods. This work addresses these issues by introducing Denoising Diffusion Restoration Models (DDRM), an efficient, unsupervised posterior sampling method. Motivated by variational inference, DDRM takes advantage of a pre-trained denoising diffusion generative model for solving any linear inverse problem. We demonstrate DDRM's versatility on several image datasets for super-resolution, deblurring, inpainting, and colorization under various amounts of measurement noise. DDRM outperforms the current leading unsupervised methods on the diverse ImageNet dataset in reconstruction quality, perceptual quality, and runtime, being  $5\times$  faster than the nearest competitor. DDRM also generalizes well for natural images out of the distribution of the observed ImageNet training set.

## WeightedSHAP: analyzing and improving Shapley based feature attributions

- Yongchan Kwon · James Zou
- abstract@[open-review](#): Shapley value is a popular approach for measuring the influence of individual features. While Shapley feature attribution is built upon desiderata from game theory, some of its constraints may be less natural in certain machine learning settings, leading to unintuitive model interpretation. In particular, the Shapley value uses the same weight for all marginal contributions---i.e. it gives the same importance when a large number of other features are given versus when a small number of other features are given. This property can be problematic if larger feature sets are more or less informative than smaller feature sets. Our work performs a rigorous analysis of the potential limitations of Shapley feature attribution. We identify simple settings where the Shapley value is mathematically suboptimal by assigning larger attributions for less influential features. Motivated by this observation, we propose WeightedSHAP, which generalizes the Shapley value and learns which marginal contributions to focus directly from data. On several real-world datasets, we demonstrate that the influential features identified by WeightedSHAP are better able to recapitulate the model's predictions compared to the features identified by the Shapley value.

## VICE: Variational Interpretable Concept Embeddings

- Lukas Muttenthaler · Charles Zheng · Patrick McClure · Robert Vandermeulen · Martin N Hebart · Francisco Pereira
- abstract@[open-review](#): A central goal in the cognitive sciences is the development of numerical models for mental representations of object concepts. This paper introduces Variational Interpretable Concept Embeddings (VICE), an approximate Bayesian method for embedding object concepts in a vector

space using data collected from humans in a triplet odd-one-out task. VICE uses variational inference to obtain sparse, non-negative representations of object concepts with uncertainty estimates for the embedding values. These estimates are used to automatically select the dimensions that best explain the data. We derive a PAC learning bound for VICE that can be used to estimate generalization performance or determine a sufficient sample size for experimental design. VICE rivals or outperforms its predecessor, SPoSE, at predicting human behavior in the triplet odd-one-out task. Furthermore, VICE's object representations are more reproducible and consistent across random initializations, highlighting the unique advantage of using VICE for deriving interpretable embeddings from human behavior.

## [You Only Live Once: Single-Life Reinforcement Learning via Learned Reward Shaping](#)

- Annie Chen · Archit Sharma · Sergey Levine · Chelsea Finn
- abstract@[open-review](#): Reinforcement learning algorithms are typically designed to learn a performant policy that can repeatedly and autonomously complete a task, typically starting from scratch. However, many real-world situations operate under a different set of assumptions: the goal might not be to learn a policy that can do the task repeatedly, but simply to perform a new task successfully once, ideally as quickly as possible, and while leveraging some prior knowledge or experience. For example, imagine a robot that is exploring another planet, where it cannot get help or supervision from humans. If it needs to navigate to a crater that it has never seen before in search of water, it does not really need to acquire a policy for reaching craters reliably, it only needs to reach this particular crater once. It must do so without the benefit of episodic resets and tackle a new, unknown terrain, but it can leverage prior experience it acquired on Earth. We formalize this problem setting, which we call single-life reinforcement learning (SLRL), where an agent must complete a task once while contending with some form of novelty in a single trial without interventions, given some prior data. In this setting, we find that algorithms designed for standard episodic reinforcement learning can struggle, as they have trouble recovering from novel states especially when informative rewards are not provided. Motivated by this observation, we also propose an algorithm, \$Q\$-weighted adversarial learning (QWALE), that addresses the dearth of supervision by employing a distribution matching strategy that leverages the agent's prior experience as guidance in novel situations. Our experiments on several single-life continuous control problems indicate that methods based on our distribution matching formulation are 20-60% more successful because they can more quickly recover from novel, out-of-distribution states.

## [SHAQ: Incorporating Shapley Value Theory into Multi-Agent Q-Learning](#)

- Jianhong Wang · Yuan Zhang · Yunjie Gu · Tae-Kyun Kim
- abstract@[open-review](#): Value factorisation is a useful technique for multi-agent reinforcement learning (MARL) in global reward game, however its underlying mechanism is not yet fully understood. This paper studies a theoretical framework for value factorisation with interpretability via Shapley value theory. We generalise Shapley value to Markov convex game called \textit{Markov Shapley value} (MSV) and apply it as a value factorisation method in global reward game, which is obtained by the equivalence between the two games. Based on the properties of MSV, we derive \textit{Shapley-Bellman optimality equation} (SBOE) to evaluate the optimal MSV, which corresponds to an optimal joint deterministic policy. Furthermore, we propose \textit{Shapley-Bellman operator} (SBO) that is proved to solve SBOE. With a stochastic approximation and some transformations, a new MARL algorithm called \textit{Shapley Q-learning} (SHAQ) is established, the implementation of which is guided by the theoretical results of SBO and MSV. We also discuss the relationship between SHAQ and relevant value factorisation methods. In the experiments SHAQ exhibits not only superior performances on all tasks but also the interpretability that agrees with the theoretical analysis.

## [Amortized Inference for Heterogeneous Reconstruction in Cryo-EM](#)

- Axel Levy · Gordon Wetzstein · Julien N.P Martel · Frederic Poitevin · Ellen Zhong
- abstract@[open-review](#): Cryo-electron microscopy (cryo-EM) is an imaging modality that provides unique insights into the dynamics of proteins and other building blocks of life. The algorithmic challenge of jointly estimating the poses, 3D structure, and conformational heterogeneity of a biomolecule from millions of noisy and randomly oriented 2D projections in a computationally efficient manner, however, remains unsolved. Our method, cryoFIRE, performs ab initio heterogeneous reconstruction with unknown poses in an amortized framework, thereby avoiding the computationally expensive step of pose search while enabling the analysis of conformational heterogeneity. Poses and conformation are jointly estimated by an encoder while a physics-based decoder aggregates the images into an implicit neural representation of the conformational space. We show that our method can provide one order of magnitude speedup on datasets containing millions of images, without any loss of accuracy. We validate that the joint estimation of poses and conformations can be amortized over the size of the dataset. For the first time, we prove that an amortized method can extract interpretable dynamic information from experimental datasets.

## [Learning from Future: A Novel Self-Training Framework for Semantic Segmentation](#)

- Ye Du · Yujun Shen · Haochen Wang · Jingjing Fei · Wei Li · Liwei Wu · Rui Zhao · Zehua Fu · Qingjie LIU
- abstract@[open-review](#): Self-training has shown great potential in semi-supervised learning. Its core idea is to use the model learned on labeled data to generate pseudo-labels for unlabeled samples, and in turn teach itself. To obtain valid supervision, active attempts typically employ a momentum teacher for pseudo-label prediction yet observe the confirmation bias issue, where the incorrect predictions may provide wrong supervision signals and get accumulated in the training process. The primary cause of such a drawback is that the prevailing self-training framework acts as guiding the current state with previous knowledge because the teacher is updated with the past student only. To alleviate this problem, we propose a novel self-training strategy, which allows the model to learn from the future. Concretely, at each training step, we first virtually optimize the student (i.e., caching the gradients without applying them to the model weights), then update the teacher with the virtual future student, and finally ask the teacher to produce pseudo-labels for the current student as the guidance. In this way, we manage to improve the quality of pseudo-labels and thus boost the performance. We also develop two variants of our future-self-training (FST) framework through peeping at the future both deeply (FST-D) and widely (FST-W). Taking the tasks of unsupervised domain adaptive semantic segmentation and semi-supervised semantic segmentation as the instances, we experimentally demonstrate the effectiveness and superiority of our approach under a wide range of settings. Code is available at <https://github.com/usr922/FST>.

## [SALSA: Attacking Lattice Cryptography with Transformers](#)

- Emily Wenger · Mingjie Chen · Francois Charton · Kristin E. Lauter
- abstract@[open-review](#): Currently deployed public-key cryptosystems will be vulnerable to attacks by full-scale quantum computers. Consequently, "quantum resistant" cryptosystems are in high demand, and lattice-based cryptosystems, based on a hard problem known as Learning With Errors (LWE), have emerged as strong contenders for standardization. In this work, we train transformers to perform modular arithmetic and mix half-trained models and statistical cryptanalysis techniques to propose SALSA: a machine learning attack on LWE-based cryptographic schemes. SALSA can fully recover secrets for small-to-mid size LWE instances with sparse binary secrets, and may scale to attack real world LWE-based cryptosystems.

## [Interpolation and Regularization for Causal Learning](#)

- Leena Chennuru Vankadara · Luca Rendsburg · Ulrike Luxburg · Debarghya Ghoshdastidar
- abstract@[open-review](#): Recent work shows that in complex model classes, interpolators can achieve statistical generalization and even be optimal for statistical learning. However, despite increasing interest in learning models with good causal properties, there is no understanding of whether such interpolators can also achieve causal generalization. To address this gap, we study causal learning from observational data through the lens of interpolation and its counterpart---regularization. Under a simple linear causal model, we derive precise asymptotics for the causal risk of the min-norm interpolator and ridge regressors in the high-dimensional regime. We find a large range of behavior that can be precisely characterized by a new measure of confounding

strength. When confounding strength is positive, which holds under independent causal mechanisms---a standard assumption in causal learning---we find that interpolators cannot be optimal. Indeed, causal learning requires stronger regularization than statistical learning. Beyond this assumption, when confounding is negative, we observe a phenomenon of self-induced regularization due to positive alignment between statistical and causal signals. Here, causal learning requires weaker regularization than statistical learning, interpolators can be optimal, and optimal regularization can even be negative.

## [Bridging the Gap: Unifying the Training and Evaluation of Neural Network Binary Classifiers](#)

- Nathan Tsoi · Kate Candon · Deyuan Li · Yofti Milkessa · Marynel Vázquez
- abstract@[open-review](#): While neural network binary classifiers are often evaluated on metrics such as Accuracy and  $F_1$ -Score, they are commonly trained with a cross-entropy objective. How can this training-evaluation gap be addressed? While specific techniques have been adopted to optimize certain confusion matrix based metrics, it is challenging or impossible in some cases to generalize the techniques to other metrics. Adversarial learning approaches have also been proposed to optimize networks via confusion matrix based metrics, but they tend to be much slower than common training methods. In this work, we propose a unifying approach to training neural network binary classifiers that combines a differentiable approximation of the Heaviside function with a probabilistic view of the typical confusion matrix values using soft sets. Our theoretical analysis shows the benefit of using our method to optimize for a given evaluation metric, such as  $F_1$ -Score, with soft sets, and our extensive experiments show the effectiveness of our approach in several domains.

## [Smooth Fictitious Play in Stochastic Games with Perturbed Payoffs and Unknown Transitions](#)

- Lucas Baudin · Rida Laraki
- abstract@[open-review](#): Recent extensions to dynamic games of the well known fictitious play learning procedure in static games were proved to globally converge to stationary Nash equilibria in two important classes of dynamic games (zero-sum and identical-interest discounted stochastic games). However, those decentralized algorithms need the players to know exactly the model (the transition probabilities and their payoffs at every stage). To overcome these strong assumptions, our paper introduces regularizations of the recent algorithms which are moreover, model-free (players don't know the transitions and their payoffs are perturbed at every stage). Our novel procedures can be interpreted as extensions to stochastic games of the classical smooth fictitious play learning procedures in static games (where players best responses are regularized, thanks to a smooth perturbation of their payoff functions). We prove the convergence of our family of procedures to stationary regularized Nash equilibria in the same classes of dynamic games (zero-sum and identical interests discounted stochastic games). The proof uses the continuous smooth best-response dynamics counterparts, and stochastic approximation methods. In the case of a MDP (a one-player stochastic game), our procedures globally converge to the optimal stationary policy of the regularized problem. In that sense, they can be seen as an alternative to the well known Q-learning procedure.

## [Uncertainty-Aware Reinforcement Learning for Risk-Sensitive Player Evaluation in Sports Game](#)

- Guiliang Liu · Yudong Luo · Oliver Schulte · Pascal Poupart
- abstract@[open-review](#): A major task of sports analytics is player evaluation. Previous methods commonly measured the impact of players' actions on desirable outcomes (e.g., goals or winning) without considering the risk induced by stochastic game dynamics. In this paper, we design an uncertainty-aware Reinforcement Learning (RL) framework to learn a risk-sensitive player evaluation metric from stochastic game dynamics. To embed the risk of a player's movements into the distribution of action-values, we model their 1) aleatoric uncertainty, which represents the intrinsic stochasticity in a sports game, and 2) epistemic uncertainty, which is due to a model's insufficient knowledge regarding Out-of-Distribution (OoD) samples. We demonstrate how a distributional Bellman operator and a feature-space density model can capture these uncertainties. Based on such uncertainty estimation, we propose a Risk-sensitive Game Impact Metric (RiGIM) that measures players' performance over a season by conditioning on a specific confidence level. Empirical evaluation, based on over 9M play-by-play ice hockey and soccer events, shows that RiGIM correlates highly with standard success measures and has a consistent risk sensitivity.

## [SCL-WC: Cross-Slide Contrastive Learning for Weakly-Supervised Whole-Slide Image Classification](#)

- Xiyue Wang · Jinxi Xiang · Jun Zhang · Sen Yang · Zhongyi Yang · Ming-Hui Wang · Jing Zhang · Wei Yang · Junzhou Huang · Xiao Han
- abstract@[open-review](#): Weakly-supervised whole-slide image (WSI) classification (WSWC) is a challenging task where a large number of unlabeled patches (instances) exist within each WSI (bag) while only a slide label is given. Despite recent progress for the multiple instance learning (MIL)-based WSI analysis, the major limitation is that it usually focuses on the easy-to-distinguish diagnosis-positive regions while ignoring positives that occupy a small ratio in the entire WSI. To obtain more discriminative features, we propose a novel weakly-supervised classification method based on cross-slide contrastive learning (called SCL-WC), which depends on task-agnostic self-supervised feature pre-extraction and task-specific weakly-supervised feature refinement and aggregation for WSI-level prediction. To enable both intra-WSI and inter-WSI information interaction, we propose a positive-negative-aware module (PNM) and a weakly-supervised cross-slide contrastive learning (WSCL) module, respectively. The WSCL aims to pull WSIs with the same disease types closer and push different WSIs away. The PNM aims to facilitate the separation of tumor-like patches and normal ones within each WSI. Extensive experiments demonstrate state-of-the-art performance of our method in three different classification tasks (e.g., over 2% of AUC in Camelyon16, 5% of F1 score in BRACS, and 3% of AUC in DiagSet). Our method also shows superior flexibility and scalability in weakly-supervised localization and semi-supervised classification experiments (e.g., first place in the BRIGHT challenge). Our code will be online.

## [Iterative Structural Inference of Directed Graphs](#)

- Aoran Wang · Jun Pang
- abstract@[open-review](#): In this paper, we propose a variational model, iterative Structural Inference of Directed Graphs (iSIDG), to infer the existence of directed interactions from observational agents' features over a time period in a dynamical system. First, the iterative process in our model feeds the learned interactions back to encourage our model to eliminate indirect interactions and to emphasize directional representation during learning. Second, we show that extra regularization terms in the objective function for smoothness, connectiveness, and sparsity prompt our model to infer a more realistic structure and to further eliminate indirect interactions. We evaluate iSIDG on various datasets including biological networks, simulated fMRI data, and physics simulations to demonstrate that our model is able to precisely infer the existence of interactions, and is significantly superior to baseline models.

## [A Geometric Perspective on Variational Autoencoders](#)

- Clément Chadebec · Stephanie Allassonnière
- abstract@[open-review](#): This paper introduces a new interpretation of the Variational Autoencoder framework by taking a fully geometric point of view. We argue that vanilla VAE models unveil naturally a Riemannian structure in their latent space and that taking into consideration those geometrical aspects can lead to better interpolations and an improved generation procedure. This new proposed sampling method consists in sampling from the uniform distribution deriving intrinsically from the learned Riemannian latent space and we show that using this scheme can make a vanilla VAE competitive and even better than more advanced versions on several benchmark datasets. Since generative models are known to be sensitive to the number of training samples we also stress the method's robustness in the low data regime.

## [Distributional Convergence of the Sliced Wasserstein Process](#)

- Jiaqi Xi · Jonathan Niles-Weed
- abstract@[open-review](#): Motivated by the statistical and computational challenges of computing Wasserstein distances in high-dimensional contexts, machine learning researchers have defined modified Wasserstein distances based on computing distances between one-dimensional projections of the measures. Different choices of how to aggregate these projected distances (averaging, random sampling, maximizing) give rise to different distances, requiring different statistical analyses. We define the *Sliced Wasserstein Process*, a stochastic process defined by the empirical Wasserstein distance between projections of empirical probability measures to all one-dimensional subspaces, and prove a uniform distributional limit theorem for this process. As a result, we obtain a unified framework in which to prove sample complexity and distributional limit results for all Wasserstein distances based on one-dimensional projections. We illustrate these results on a number of examples where no distributional limits were previously known.

## [Exploration With a Finite Brain](#)

- Marcel Binz · Eric Schulz
- abstract@[open-review](#): Equipping artificial agents with useful exploration mechanisms remains a challenge to this day. Humans, on the other hand, seem to manage the trade-off between exploration and exploitation effortlessly. In the present article, we put forward the hypothesis that they accomplish this by making optimal use of limited computational resources. We study this hypothesis by meta-learning reinforcement learning algorithms that sacrifice performance for a shorter description length (defined as the number of bits required to implement the given algorithm). The emerging class of models captures human exploration behavior better than previously considered approaches, such as Boltzmann exploration, upper confidence bound algorithms, and Thompson sampling. We additionally demonstrate that changing the description length in our class of models produces the intended effects: reducing description length captures the behavior of brain-lesioned patients while increasing it mirrors cognitive development during adolescence.

## [A fully adaptive trust-region method](#)

- Fadi Hamad · Oliver Hinder
- abstract@[open-review](#): Adaptive second-order methods, such as Cartis, Gould, and Toint's adaptive cubic regularized Newton method (ARC), attempt to maintain strong convergence guarantees without depending on conservative estimates of problem properties such as Lipschitz constants. However, on close inspection, one can show existing 'adaptive methods' have theoretical guarantees with severely suboptimal dependence on problem properties such as the Lipschitz constant of the Hessian. For example, Cartis, Gould, and Toint's bound the number of iterations of their algorithm by  $\mathcal{O}(\Delta_f L^{3/2} \sigma_{\min}^{-1} \epsilon^{-3/2})$  where  $\sigma_{\min}$  is the smallest observed regularization parameter. However, as  $\sigma_{\min}$  could be arbitrarily small, this bound can be vacuous. Unfortunately, subsequent adaptive second-order methods are built on their ideas, and therefore also suffer from this issue. We present the first adaptive second-order method which circumvents this issue and requires at most  $\mathcal{O}(\Delta_f L^{1/2} \epsilon^{-3/2}) + \tilde{\mathcal{O}}(1)$  iterations to find an  $\epsilon$ -approximate stationary point, matching the optimal iteration bound up to an additive logarithmic term. Our method is a simple variant of a classic trust-region method and in our experiments performs competitively with both ARC and a classical trust-region method.

## [Inductive Logical Query Answering in Knowledge Graphs](#)

- Mikhail Galkin · Zhaocheng Zhu · Hongyu Ren · Jian Tang
- abstract@[open-review](#): Formulating and answering logical queries is a standard communication interface for knowledge graphs (KGs) and their representations. Alleviating the notorious incompleteness of real-world KGs, neural methods achieved impressive results in link prediction and complex query answering tasks by learning representations of entities, relations, and queries. Still, most existing query answering methods are inherently transductive and cannot be generalized to KGs containing new entities without retraining entity embeddings. In this work, we study the inductive query answering task where inference is performed on a graph containing new entities with queries over both seen and unseen entities. To this end, we devise two mechanisms leveraging inductive node and relational structure representations powered by graph neural networks (GNNs). Experimentally, we show that inductive models are able to perform logical reasoning at inference time over unseen nodes generalizing to graphs up to 500% larger than training ones. Exploring the efficiency--effectiveness trade-off, we find the inductive relational structure method generally achieves higher performance, while the inductive node representation method is able to answer complex queries in the inference-only regime without any training on queries and scale to graphs of millions of nodes.

## [Posterior Matching for Arbitrary Conditioning](#)

- Ryan Strauss · Junier B Oliva
- abstract@[open-review](#): Arbitrary conditioning is an important problem in unsupervised learning, where we seek to model the conditional densities  $p(\mathbf{x}_u | \mathbf{x}_o)$  that underly some data, for all possible non-intersecting subsets  $\mathbf{o}, u \subset \{1, \dots, d\}$ . However, the vast majority of density estimation only focuses on modeling the joint distribution  $p(\mathbf{x})$ , in which important conditional dependencies between features are opaque. We propose a simple and general framework, coined Posterior Matching, that enables Variational Autoencoders (VAEs) to perform arbitrary conditioning, without modification to the VAE itself. Posterior Matching applies to the numerous existing VAE-based approaches to joint density estimation, thereby circumventing the specialized models required by previous approaches to arbitrary conditioning. We find that Posterior Matching is comparable or superior to current state-of-the-art methods for a variety of tasks with an assortment of VAEs (e.g. discrete, hierarchical, VaDE).

## [TransBoost: Improving the Best ImageNet Performance using Deep Transduction](#)

- Omer Belhasin · Guy Bar-Shalom · Ran El-Yaniv
- abstract@[open-review](#): This paper deals with deep transductive learning, and proposes TransBoost as a procedure for fine-tuning any deep neural model to improve its performance on any (unlabeled) test set provided at training time. TransBoost is inspired by a large margin principle and is efficient and simple to use. The ImageNet classification performance is consistently and significantly improved with TransBoost on many architectures such as ResNets, MobileNetV3-L, EfficientNetB0, ViT-S, and ConvNext-T. Additionally we show that TransBoost is effective on a wide variety of image classification datasets.

## [Hardness in Markov Decision Processes: Theory and Practice](#)

- Michelangelo Conserva · Paulo Rauher
- abstract@[open-review](#): Meticulously analysing the empirical strengths and weaknesses of reinforcement learning methods in hard (challenging) environments is essential to inspire innovations and assess progress in the field. In tabular reinforcement learning, there is no well-established standard selection of environments to conduct such analysis, which is partially due to the lack of a widespread understanding of the rich theory of hardness of environments. The goal of this paper is to unlock the practical usefulness of this theory through four main contributions. First, we present a systematic survey of the theory of hardness, which also identifies promising research directions. Second, we introduce "Colosseum", a pioneering Python package that enables empirical hardness analysis and implements a principled benchmark composed of environments that are diverse with respect to different measures of hardness. Third, we present an empirical comparison that provides new insights into current (efficiently computable) measures. Finally, we report the results of state-of-the-art tabular reinforcement learning algorithms in our newly proposed benchmark. Our contributions to tabular reinforcement learning are intended as solid steps towards the development of more principled benchmarks for the non-tabular setting.

## [Flamingo: a Visual Language Model for Few-Shot Learning](#)

- Jean-Baptiste Alayrac · Jeff Donahue · Pauline Luc · Antoine Miech · Iain Barr · Yana Hasson · Karel Lenc · Arthur Mensch · Katherine Millican · Malcolm Reynolds · Roman Ring · Eliza Rutherford · Serkan Cabi · Tengda Han · Zhitao Gong · Sina Samangooei · Marianne Monteiro · Jacob L Menick · Sebastian Borgeaud · Andy Brock · Aida Nematzadeh · Sahand Sharifzadeh · Mikołaj Bikowski · Ricardo Barreira · Oriol Vinyals · Andrew Zisserman · Karen Simonyan
- abstract@[open-review](#): Building models that can be rapidly adapted to novel tasks using only a handful of annotated examples is an open challenge for multimodal machine learning research. We introduce Flamingo, a family of Visual Language Models (VLM) with this ability. We propose key architectural innovations to: (i) bridge powerful pretrained vision-only and language-only models, (ii) handle sequences of arbitrarily interleaved visual and textual data, and (iii) seamlessly ingest images or videos as inputs. Thanks to their flexibility, Flamingo models can be trained on large-scale multimodal web corpora containing arbitrarily interleaved text and images, which is key to endow them with in-context few-shot learning capabilities. We perform a thorough evaluation of our models, exploring and measuring their ability to rapidly adapt to a variety of image and video tasks. These include open-ended tasks such as visual question-answering, where the model is prompted with a question which it has to answer, captioning tasks, which evaluate the ability to describe a scene or an event, and close-ended tasks such as multiple-choice visual question-answering. For tasks lying anywhere on this spectrum, a single Flamingo model can achieve a new state of the art with few-shot learning, simply by prompting the model with task-specific examples. On numerous benchmarks, Flamingo outperforms models fine-tuned on thousands of times more task-specific data.

## [Using natural language and program abstractions to instill human inductive biases in machines](#)

- Sreejan Kumar · Carlos G. Correa · Ishita Dasgupta · Raja Marjieh · Michael Y Hu · Robert Hawkins · Jonathan D Cohen · nathaniel daw · Karthik Narasimhan · Tom Griffiths
- abstract@[open-review](#): Strong inductive biases give humans the ability to quickly learn a variety of tasks. Although meta-learning is a method to endow neural networks with useful inductive biases, agents trained by meta-learning may sometimes acquire very different strategies from humans. We show that co-training these agents on predicting representations from natural language task descriptions and programs induced to generate such tasks guides them toward human-like inductive biases. Human-generated language descriptions and program induction models that add new learned primitives both contain abstract concepts that can compress description length. Co-training on these representations result in more human-like behavior in downstream meta-reinforcement learning agents than less abstract controls (synthetic language descriptions, program induction without learned primitives), suggesting that the abstraction supported by these representations is key.

## [Tracking Functional Changes in Nonstationary Signals with Evolutionary Ensemble Bayesian Model for Robust Neural Decoding](#)

- Xinyun Zhu · Yu Qi · Gang Pan · Yueming Wang
- abstract@[open-review](#): Neural signals are typical nonstationary data where the functional mapping between neural activities and the intentions (such as the velocity of movements) can occasionally change. Existing studies mostly use a fixed neural decoder, thus suffering from an unstable performance given neural functional changes. We propose a novel evolutionary ensemble framework (EvoEnsemble) to dynamically cope with changes in neural signals by evolving the decoder model accordingly. EvoEnsemble integrates evolutionary computation algorithms in a Bayesian framework where the fitness of models can be sequentially computed with their likelihoods according to the incoming data at each time slot, which enables online tracking of time-varying functions. Two strategies of evolve-at-changes and history-model-archive are designed to further improve efficiency and stability. Experiments with simulations and neural signals demonstrate that EvoEnsemble can track the changes in functions effectively thus improving the accuracy and robustness of neural decoding. The improvement is most significant in neural signals with functional changes.

## [Rethinking the compositionality of point clouds through regularization in the hyperbolic space](#)

- Antonio Montanaro · Diego Valsesia · Enrico Magli
- abstract@[open-review](#): Point clouds of 3D objects exhibit an inherent compositional nature where simple parts can be assembled into progressively more complex shapes to form whole objects. Explicitly capturing such part-whole hierarchy is a long-sought objective in order to build effective models, but its tree-like nature has made the task elusive. In this paper, we propose to embed the features of a point cloud classifier into the hyperbolic space and explicitly regularize the space to account for the part-whole hierarchy. The hyperbolic space is the only space that can successfully embed the tree-like nature of the hierarchy. This leads to substantial improvements in the performance of state-of-art supervised models for point cloud classification.

## [Data-Efficient Structured Pruning via Submodular Optimization](#)

- Marwa El Halabi · Suraj Srinivas · Simon Lacoste-Julien
- abstract@[open-review](#): Structured pruning is an effective approach for compressing large pre-trained neural networks without significantly affecting their performance. However, most current structured pruning methods do not provide any performance guarantees, and often require fine-tuning, which makes them inapplicable in the limited-data regime. We propose a principled data-efficient structured pruning method based on submodular optimization. In particular, for a given layer, we select neurons/channels to prune and corresponding new weights for the next layer, that minimize the change in the next layer's input induced by pruning. We show that this selection problem is a weakly submodular maximization problem, thus it can be provably approximated using an efficient greedy algorithm. Our method is guaranteed to have an exponentially decreasing error between the original model and the pruned model outputs w.r.t the pruned size, under reasonable assumptions. It is also one of the few methods in the literature that uses only a limited-number of training data and no labels. Our experimental results demonstrate that our method outperforms state-of-the-art methods in the limited-data regime.

## [Online Convex Optimization with Hard Constraints: Towards the Best of Two Worlds and Beyond](#)

- Hengquan Guo · Xin Liu · Honghao Wei · Lei Ying
- abstract@[open-review](#): This paper considers online convex optimization with hard constraints and analyzes achievable regret and cumulative hard constraint violation (violation for short). The problem distinguishes itself from online convex optimization with soft constraints, where a violation at one round can be compensated/cancelled by a conservative decision at a different round. We propose a RECtified Online Optimization algorithm (RECOO) and consider two settings: fixed constraints and adversarial constraints. Both settings have been considered in the literature. Compared with existing results, RECOO achieves the best of two worlds and beyond. For the fixed-constraints setting, RECOO achieves  $\mathcal{O}(\sqrt{T})$  regret and  $\mathcal{O}(1)$  violation, where  $T$  is the learning horizon. The best known results in this case are  $\mathcal{O}(\sqrt{T})$  regret and  $\mathcal{O}(T^{1/4})$  violation. For the adversarial-constraints setting, it guarantees  $\mathcal{O}(\sqrt{T})$  regret and  $\mathcal{O}(T^{3/4})$  violation, which match the best existing results. When the loss functions are strongly convex, RECOO can guarantee  $\mathcal{O}(\log T)$  regret and  $\mathcal{O}(1)$  violation for fixed constraints, and  $\mathcal{O}(\log T)$  regret and  $\mathcal{O}(\sqrt{T \log T})$  violation for adversarial constraints. Both these results are order-wise better than the existing bounds. The regret and violation bounds mentioned above use the best fixed decision in hindsight as the baseline. This paper further considers a dynamic baseline where the comparator sequence is time-varying. This paper shows that RECOO not only improves the existing results in the fixed-constraints setting but also for the first time, guarantees dynamic regret and violation bounds in the adversarial-constraints setting. Our experiment results confirm that RECOO outperforms several existing algorithms for both fixed and adversarial constraints.

## [What You See is What You Classify: Black Box Attributions](#)

- Steven Stalder · Nathanael Perraudin · Radhakrishna Achanta · Fernando Perez-Cruz · Michele Volpi
- abstract@[open-review](#): An important step towards explaining deep image classifiers lies in the identification of image regions that contribute to individual class scores in the model's output. However, doing this accurately is a difficult task due to the black-box nature of such networks. Most existing approaches find such attributions either using activations and gradients or by repeatedly perturbing the input. We instead address this challenge by training a second deep network, the Explainer, to predict attributions for a pre-trained black-box classifier, the Explanandum. These attributions are in the form of masks that only show the classifier-relevant parts of an image, masking out the rest. Our approach produces sharper and more boundary-precise masks when compared to the saliency maps generated by other methods. Moreover, unlike most existing approaches, ours is capable of directly generating very distinct class-specific masks. Finally, the proposed method is very efficient for inference since it only takes a single forward pass through the Explainer to generate all class-specific masks. We show that our attributions are superior to established methods both visually and quantitatively, by evaluating them on the PASCAL VOC-2007 and Microsoft COCO-2014 datasets.

## [Deep Attentive Belief Propagation: Integrating Reasoning and Learning for Solving Constraint Optimization Problems](#)

- Yanchen Deng · Shufeng Kong · Caihua Liu · Bo An
- abstract@[open-review](#): Belief Propagation (BP) is an important message-passing algorithm for various reasoning tasks over graphical models, including solving the Constraint Optimization Problems (COPs). It has been shown that BP can achieve state-of-the-art performance on various benchmarks by mixing old and new messages before sending the new one, i.e., damping. However, existing methods on tuning a static damping factor for BP not only is laborious but also harms their performance. Moreover, existing BP algorithms treat each variable node's neighbors equally when composing a new message, which also limits their exploration ability. To address these issues, we seamlessly integrate BP, Gated Recurrent Units (GRUs), and Graph Attention Networks (GATs) within the message-passing framework to reason about dynamic weights and damping factors for composing new BP messages. Our model, Deep Attentive Belief Propagation (DABP), takes the factor graph and the BP messages in each iteration as the input and infers the optimal weights and damping factors through GRUs and GATs, followed by a multi-head attention layer. Furthermore, unlike existing neural-based BP variants, we propose a novel self-supervised learning algorithm for DABP with a smoothed solution cost, which does not require expensive training labels and also avoids the common out-of-distribution issue through efficient online learning. Extensive experiments show that our model significantly outperforms state-of-the-art baselines.

## [Cost-efficient Gaussian tensor network embeddings for tensor-structured inputs](#)

- Linjian Ma · Edgar Solomonik
- abstract@[open-review](#): This work discusses tensor network embeddings, which are random matrices (\$S\$) with tensor network structure. These embeddings have been used to perform dimensionality reduction of tensor network structured inputs \$x\$ and accelerate applications such as tensor decomposition and kernel regression. Existing works have designed embeddings for inputs \$x\$ with specific structures, such as the Kronecker product or Khatri-Rao product, such that the computational cost for calculating \$Sx\$ is efficient. We provide a systematic way to design tensor network embeddings consisting of Gaussian random tensors, such that for inputs with more general tensor network structures, both the sketch size (row size of \$S\$) and the sketching computational cost are low. We analyze general tensor network embeddings that can be reduced to a sequence of sketching matrices. We provide a sufficient condition to quantify the accuracy of such embeddings and derive sketching asymptotic cost lower bounds using embeddings that satisfy this condition and have a sketch size lower than any input dimension. We then provide an algorithm to efficiently sketch input data using such embeddings. The sketch size of the embedding used in the algorithm has a linear dependence on the number of sketching dimensions of the input. Assuming tensor contractions are performed with classical dense matrix multiplication algorithms, this algorithm achieves asymptotic cost within a factor of \$O(\sqrt{m})\$ of our cost lower bound, where \$m\$ is the sketch size. Further, when each tensor in the input has a dimension that needs to be sketched, this algorithm yields the optimal sketching asymptotic cost. We apply our sketching analysis to inexact tensor decomposition optimization algorithms. We provide a sketching algorithm for CP decomposition that is asymptotically faster than existing work in multiple regimes, and show optimality of an existing algorithm for tensor train rounding.

## [Beyond accuracy: generalization properties of bio-plausible temporal credit assignment rules](#)

- Yuhan Helena Liu · Arna Ghosh · Blake Richards · Eric Shea-Brown · Guillaume Lajoie
- abstract@[open-review](#): To unveil how the brain learns, ongoing work seeks biologically-plausible approximations of gradient descent algorithms for training recurrent neural networks (RNNs). Yet, beyond task accuracy, it is unclear if such learning rules converge to solutions that exhibit different levels of generalization than their non-biologically-plausible counterparts. Leveraging results from deep learning theory based on loss landscape curvature, we ask: how do biologically-plausible gradient approximations affect generalization? We first demonstrate that state-of-the-art biologically-plausible learning rules for training RNNs exhibit worse and more variable generalization performance compared to their machine learning counterparts that follow the true gradient more closely. Next, we verify that such generalization performance is correlated significantly with loss landscape curvature, and we show that biologically-plausible learning rules tend to approach high-curvature regions in synaptic weight space. Using tools from dynamical systems, we derive theoretical arguments and present a theorem explaining this phenomenon. This predicts our numerical results, and explains why biologically-plausible rules lead to worse and more variable generalization properties. Finally, we suggest potential remedies that could be used by the brain to mitigate this effect. To our knowledge, our analysis is the first to identify the reason for this generalization gap between artificial and biologically-plausible learning rules, which can help guide future investigations into how the brain learns solutions that generalize.

## [BagFlip: A Certified Defense Against Data Poisoning](#)

- Yuhao Zhang · Aws Albarghouthi · Loris D'Antoni
- abstract@[open-review](#): Machine learning models are vulnerable to data-poisoning attacks, in which an attacker maliciously modifies the training set to change the prediction of a learned model. In a trigger-less attack, the attacker can modify the training set but not the test inputs, while in a backdoor attack the attacker can also modify test inputs. Existing model-agnostic defense approaches either cannot handle backdoor attacks or do not provide effective certificates (i.e., a proof of a defense). We present BagFlip, a model-agnostic certified approach that can effectively defend against both trigger-less and backdoor attacks. We evaluate BagFlip on image classification and malware detection datasets. BagFlip is equal to or more effective than the state-of-the-art approaches for trigger-less attacks and more effective than the state-of-the-art approaches for backdoor attacks.

## [Discovery of Single Independent Latent Variable](#)

- Uri Shaham · Jonathan Svirsky · Ori Katz · Ronen Talmon
- abstract@[open-review](#): Latent variable discovery is a central problem in data analysis with a broad range of applications in applied science. In this work, we consider data given as an invertible mixture of two statistically independent components, and assume that one of the components is observed while the other is hidden. Our goal is to recover the hidden component. For this purpose, we propose an autoencoder equipped with a discriminator. Unlike the standard nonlinear ICA problem, which was shown to be non-identifiable, in the special case of ICA we consider here, we show that our approach can recover the component of interest up to entropy-preserving transformation. We demonstrate the performance of the proposed approach on several datasets, including image synthesis, voice cloning, and fetal ECG extraction.

## [GAUDI: A Neural Architect for Immersive 3D Scene Generation](#)

- Miguel Angel Bautista · Pengsheng Guo · Samira Abnar · Walter Talbott · Alexander Toshev · Zhiyuan Chen · Laurent Dinh · Shuangfei Zhai · Hanlin Goh · Daniel Ulbricht · Afshin Dehghan · Joshua Susskind
- abstract@[open-review](#): We introduce GAUDI, a generative model capable of capturing the distribution of complex and realistic 3D scenes that can be rendered immersively from a moving camera. We tackle this challenging problem with a scalable yet powerful approach, where we first optimize a latent representation that disentangles radiance fields and camera poses. This latent representation is then used to learn a generative model that enables both unconditional and conditional generation of 3D scenes. Our model generalizes previous works that focus on single objects by removing the assumption that the camera pose distribution can be shared across samples. We show that GAUDI obtains state-of-the-art performance in the unconditional generative setting across multiple datasets and allows for conditional generation of 3D scenes given conditioning variables like sparse image observations or text that describes the scene.

## [Between Stochastic and Adversarial Online Convex Optimization: Improved Regret Bounds via Smoothness](#)

- Sarah Sachs · Hedi Hadji · Tim van Erven · Cristóbal Guzmán
- abstract@[open-review](#): Stochastic and adversarial data are two widely studied settings in online learning. But many optimization tasks are neither i.i.d. nor fully adversarial, which makes it of fundamental interest to get a better theoretical understanding of the world between these extremes. In this work we establish novel regret bounds for online convex optimization in a setting that interpolates between stochastic i.i.d. and fully adversarial losses. By exploiting smoothness of the expected losses, these bounds replace a dependence on the maximum gradient length by the variance of the gradients, which was previously known only for linear losses. In addition, they weaken the i.i.d.\ assumption by allowing, for example, adversarially poisoned rounds, which were previously considered in the expert and bandit setting. Our results extend this to the online convex optimization framework. In the fully i.i.d.\ case, our bounds match the rates one would expect from results in stochastic acceleration, and in the fully adversarial case they gracefully deteriorate to match the minimax regret.%We further provide lower bounds showing that our regret upper bounds are%tight for all intermediate regimes in terms of the cumulative stochastic variance and the adversarial variation. We further provide lower bounds showing that our regret upper bounds are tight for all intermediate regimes in terms of the stochastic variance and the adversarial variation of the loss gradients.

## [MoGDE: Boosting Mobile Monocular 3D Object Detection with Ground Depth Estimation](#)

- Yunsong Zhou · Quan Liu · Hongzi Zhu · Yunzhe Li · Shan Chang · Minyi Guo
- abstract@[open-review](#): Monocular 3D object detection (Mono3D) in mobile settings (e.g., on a vehicle, a drone, or a robot) is an important yet challenging task. Due to the near-far disparity phenomenon of monocular vision and the ever-changing camera pose, it is hard to acquire high detection accuracy, especially for far objects. Inspired by the insight that the depth of an object can be well determined according to the depth of the ground where it stands, in this paper, we propose a novel Mono3D framework, called MoGDE, which constantly estimates the corresponding ground depth of an image and then utilizes the estimated ground depth information to guide Mono3D. To this end, we utilize a pose detection network to estimate the pose of the camera and then construct a feature map portraying pixel-level ground depth according to the 3D-to-2D perspective geometry. Moreover, to improve Mono3D with the estimated ground depth, we design an RGB-D feature fusion network based on the transformer structure, where the long-range self-attention mechanism is utilized to effectively identify ground-contacting points and pin the corresponding ground depth to the image feature map. We conduct extensive experiments on the real-world KITTI dataset. The results demonstrate that MoGDE can effectively improve the Mono3D accuracy and robustness for both near and far objects. MoGDE yields the best performance compared with the state-of-the-art methods by a large margin and is ranked number one on the KITTI 3D benchmark.

## [Semi-supervised Semantic Segmentation with Prototype-based Consistency Regularization](#)

- Haiming Xu · Lingqiao Liu · Qiuchen Bian · Zhen Yang
- abstract@[open-review](#): Semi-supervised semantic segmentation requires the model to effectively propagate the label information from limited annotated images to unlabeled ones. A challenge for such a per-pixel prediction task is the large intra-class variation, i.e., regions belonging to the same class may exhibit a very different appearance even in the same picture. This diversity will make the label propagation hard from pixels to pixels. To address this problem, we propose a novel approach to regularize the distribution of within-class features to ease label propagation difficulty. Specifically, our approach encourages the consistency between the prediction from a linear predictor and the output from a prototype-based predictor, which implicitly encourages features from the same pseudo-class to be close to at least one within-class prototype while staying far from the other between-class prototypes. By further incorporating CutMix operations and a carefully-designed prototype maintenance strategy, we create a semi-supervised semantic segmentation algorithm that demonstrates superior performance over the state-of-the-art methods from extensive experimental evaluation on both Pascal VOC and Cityscapes benchmarks.

## [Active Surrogate Estimators: An Active Learning Approach to Label-Efficient Model Evaluation](#)

- Jannik Kossen · Sebastian Farquhar · Yarin Gal · Thomas Rainforth
- abstract@[open-review](#): We propose Active Surrogate Estimators (ASEs), a new method for label-efficient model evaluation. Evaluating model performance is a challenging and important problem when labels are expensive. ASEs address this active testing problem using a surrogate-based estimation approach that interpolates the errors of points with unknown labels, rather than forming a Monte Carlo estimator. ASEs actively learn the underlying surrogate, and we propose a novel acquisition strategy, XWED, that tailors this learning to the final estimation task. We find that ASEs offer greater label-efficiency than the current state-of-the-art when applied to challenging model evaluation problems for deep neural networks.

## [Universality of group convolutional neural networks based on ridgelet analysis on groups](#)

- Sho Sonoda · Isao Ishikawa · Masahiro Ikeda
- abstract@[open-review](#): We investigate the approximation property of group convolutional neural networks (GCNNs) based on the ridgelet theory. We regard a group convolution as a matrix element of a group representation, and formulate a versatile GCNN as a nonlinear mapping between group representations, which covers typical GCNN literatures such as a cyclic convolution on a multi-channel image, permutation-invariant datasets (Deep Sets), and  $\mathbb{E}(n)$ -equivariant convolutions. The ridgelet transform is an analysis operator of a depth-2 network, namely, it maps an arbitrary given target function  $f$  to the weight  $\gamma$  of a network  $S[\gamma]$  so that the network represents the function as  $S[\gamma]=f$ . It has been known only for fully-connected networks, and this study is the first to present the ridgelet transform for (G)CNNs. Since the ridgelet transform is given as a closed-form integral operator, it provides a constructive proof of the  $\mathcal{CC}$ -universality of GCNNs. Unlike previous universality arguments on CNNs, we do not need to convert/modify the networks into other universal approximators such as invariant polynomials and fully-connected networks.

## [Federated Submodel Optimization for Hot and Cold Data Features](#)

- Yucheng Ding · Chaoyue Niu · Fan Wu · Shaojie Tang · Chengfei Lyu · yanghe feng · Guihai Chen
- abstract@[open-review](#): We focus on federated learning in practical recommender systems and natural language processing scenarios. The global model for federated optimization typically contains a large and sparse embedding layer, while each client's local data tend to interact with part of features, updating only a small submodel with the feature-related embedding vectors. We identify a new and important issue that distinct data features normally involve different numbers of clients, generating the differentiation of hot and cold features. We further reveal that the classical federated averaging algorithm (FedAvg) or its variants, which randomly selects clients to participate and uniformly averages their submodel updates, will be severely slowed down, because different parameters of the global model are optimized at different speeds. More specifically, the model parameters related to hot (resp.,

cold) features will be updated quickly (resp., slowly). We thus propose federated submodel averaging (FedSubAvg), which introduces the number of feature-related clients as the metric of feature heat to correct the aggregation of submodel updates. We prove that due to the dispersion of feature heat, the global objective is ill-conditioned, and FedSubAvg works as a suitable diagonal preconditioner. We also rigorously analyze FedSubAvg's convergence rate to stationary points. We finally evaluate FedSubAvg over several public and industrial datasets. The evaluation results demonstrate that FedSubAvg significantly outperforms FedAvg and its variants.

## [Towards Reliable Simulation-Based Inference with Balanced Neural Ratio Estimation](#)

- Arnaud Delaunoy · Joeri Hermans · François Rozet · Antoine Wehenkel · Gilles Louppe
- abstract@[open-review](#): Modern approaches for simulation-based inference build upon deep learning surrogates to enable approximate Bayesian inference with computer simulators. In practice, the estimated posteriors' computational faithfulness is, however, rarely guaranteed. For example, Hermans et al., 2021 have shown that current simulation-based inference algorithms can produce posteriors that are overconfident, hence risking false inferences. In this work, we introduce Balanced Neural Ratio Estimation (BNRE), a variation of the NRE algorithm designed to produce posterior approximations that tend to be more conservative, hence improving their reliability, while sharing the same Bayes optimal solution. We achieve this by enforcing a balancing condition that increases the quantified uncertainty in low simulation budget regimes while still converging to the exact posterior as the budget increases. We provide theoretical arguments showing that BNRE tends to produce posterior surrogates that are more conservative than NRE's. We evaluate BNRE on a wide variety of tasks and show that it produces conservative posterior surrogates on all tested benchmarks and simulation budgets. Finally, we emphasize that BNRE is straightforward to implement over NRE and does not introduce any computational overhead.

## [Gold-standard solutions to the Schrödinger equation using deep learning: How much physics do we need?](#)

- Leon Gerard · Michael Scherbela · Philipp Marquetand · Philipp Grohs
- abstract@[open-review](#): Finding accurate solutions to the Schrödinger equation is the key unsolved challenge of computational chemistry. Given its importance for the development of new chemical compounds, decades of research have been dedicated to this problem, but due to the large dimensionality even the best available methods do not yet reach the desired accuracy. Recently the combination of deep learning with Monte Carlo methods has emerged as a promising way to obtain highly accurate energies and moderate scaling of computational cost. In this paper we significantly contribute towards this goal by introducing a novel deep-learning architecture that achieves 40-70% lower energy error at 6x lower computational cost compared to previous approaches. Using our method we establish a new benchmark by calculating the most accurate variational ground state energies ever published for a number of different atoms and molecules. We systematically break down and measure our improvements, focusing in particular on the effect of increasing physical prior knowledge. We surprisingly find that increasing the prior knowledge given to the architecture can actually decrease accuracy.

## [Deconfounded Representation Similarity for Comparison of Neural Networks](#)

- Tianyu Cui · Yogesh Kumar · Pekka Marttinen · Samuel Kaski
- abstract@[open-review](#): Similarity metrics such as representational similarity analysis (RSA) and centered kernel alignment (CKA) have been used to understand neural networks by comparing their layer-wise representations. However, these metrics are confounded by the population structure of data items in the input space, leading to inconsistent conclusions about the functional similarity between neural networks, such as spuriously high similarity of completely random neural networks and inconsistent domain relations in transfer learning. We introduce a simple and generally applicable fix to adjust for the confounder with covariate adjustment regression, which improves the ability of CKA and RSA to reveal functional similarity and also retains the intuitive invariance properties of the original similarity measures. We show that deconfounding the similarity metrics increases the resolution of detecting functionally similar neural networks across domains. Moreover, in real-world applications, deconfounding improves the consistency between CKA and domain similarity in transfer learning, and increases the correlation between CKA and model out-of-distribution accuracy similarity.

## [Universally Expressive Communication in Multi-Agent Reinforcement Learning](#)

- Matthew Morris · Thomas D Barrett · Arnu Pretorius
- abstract@[open-review](#): Allowing agents to share information through communication is crucial for solving complex tasks in multi-agent reinforcement learning. In this work, we consider the question of whether a given communication protocol can express an arbitrary policy. By observing that many existing protocols can be viewed as instances of graph neural networks (GNNs), we demonstrate the equivalence of joint action selection to node labelling. With standard GNN approaches provably limited in their expressive capacity, we draw from existing GNN literature and consider augmenting agent observations with: (1) unique agent IDs and (2) random noise. We provide a theoretical analysis as to how these approaches yield universally expressive communication, and also prove them capable of targeting arbitrary sets of actions for identical agents. Empirically, these augmentations are found to improve performance on tasks where expressive communication is required, whilst, in general, the optimal communication protocol is found to be task-dependent.

## [Frank-Wolfe-based Algorithms for Approximating Tyler's M-estimator](#)

- Dan Garber · Lior Danon
- abstract@[open-review](#): Tyler's M-estimator is a well known procedure for robust and heavy-tailed covariance estimation. Tyler himself suggested an iterative fixed-point algorithm for computing his estimator however, it requires super-linear (in the size of the data) runtime per iteration, which maybe prohibitive in large scale. In this work we propose, to the best of our knowledge, the first Frank-Wolfe-based algorithms for computing Tyler's estimator. One variant uses standard Frank-Wolfe steps, the second also considers away-steps (AFW), and the third is a geodesic version of AFW (GAFW). AFW provably requires, up to a log factor, only linear time per iteration, while GAFW runs in linear time (up to a log factor) in a large  $n$  (number of data-points) regime. All three variants are shown to provably converge to the optimal solution with sublinear rate, under standard assumptions, despite the fact that the underlying optimization problem is not convex nor smooth. Under an additional fairly mild assumption, that holds with probability 1 when the (normalized) data-points are i.i.d. samples from a continuous distribution supported on the entire unit sphere, AFW and GAFW are proved to converge with linear rates. Importantly, all three variants are parameter-free and use adaptive step-sizes.

## [Fuzzy Learning Machine](#)

- Junbiao Cui · Jiye Liang
- abstract@[open-review](#): Classification is one of the most important problems in machine learning and the nature of it is concept cognition. So far, dozens of different classifiers have been designed. Although their working mechanisms vary widely, few of them fully consider concept cognition. In this paper, a new learning machine, fuzzy learning machine (FLM), is proposed from the perspective of concept cognition. Inspired by cognitive science, its working mechanism is of strong interpretability. At the same time, FLM roots in set theory and fuzzy set theory, so FLM has a solid mathematical foundation. The systematic experimental results on a large number of data sets show that FLM can achieve excellent performance, even with the simple implementation.

## [Attracting and Dispersing: A Simple Approach for Source-free Domain Adaptation](#)

- Shiqi Yang · Yaxing Wang · Kai Wang · Shangling Jui · Joost van de Weijer

- abstract@[open-review](#): We propose a simple but effective source-free domain adaptation (SFDA) method. Treating SFDA as an unsupervised clustering problem and following the intuition that local neighbors in feature space should have more similar predictions than other features, we propose to optimize an objective of prediction consistency. This objective encourages local neighborhood features in feature space to have similar predictions while features farther away in feature space have dissimilar predictions, leading to efficient feature clustering and cluster assignment simultaneously. For efficient training, we seek to optimize an upper-bound of the objective resulting in two simple terms. Furthermore, we relate popular existing methods in domain adaptation, source-free domain adaptation and contrastive learning via the perspective of discriminability and diversity. The experimental results prove the superiority of our method, and our method can be adopted as a simple but strong baseline for future research in SFDA. Our method can be also adapted to source-free open-set and partial-set DA which further shows the generalization ability of our method. Code is available in [https://github.com/Albert0147/AaD\\_SFDA](https://github.com/Albert0147/AaD_SFDA).

## [Mingling Foresight with Imagination: Model-Based Cooperative Multi-Agent Reinforcement Learning](#)

- Zhiwei Xu · dapeng li · Bin Zhang · Yuan Zhan · Yunpeng Baiia · Guoliang Fan
- abstract@[open-review](#): Recently, model-based agents have achieved better performance than model-free ones using the same computational budget and training time in single-agent environments. However, due to the complexity of multi-agent systems, it is tough to learn the model of the environment. The significant compounding error may hinder the learning process when model-based methods are applied to multi-agent tasks. This paper proposes an implicit model-based multi-agent reinforcement learning method based on value decomposition methods. Under this method, agents can interact with the learned virtual environment and evaluate the current state value according to imagined future states in the latent space, making agents have the foresight. Our approach can be applied to any multi-agent value decomposition method. The experimental results show that our method improves the sample efficiency in different partially observable Markov decision process domains.

## [Generalization Analysis on Learning with a Concurrent Verifier](#)

- Masaaki Nishino · Kengo Nakamura · Norihito Yasuda
- abstract@[open-review](#): Machine learning technologies have been used in a wide range of practical systems. In practical situations, it would be natural to expect input-output pairs of a machine learning model to satisfy some requirements. However, it is hard to obtain a model satisfying the requirements by just learning from examples. A simple solution is to add a module that checks whether the input-output pairs meet the requirements and then modifies the model's outputs. Such a module, we call {em concurrent verifier}, can give a certification, but how the generalizability of the machine learning model changes by using a concurrent verifier is unclear. This paper gives a generalization analysis of learning with a concurrent verifier. We analyze how the learnability of a machine learning model changes when we use a concurrent verifier, and show the condition where we can obtain a guaranteed hypothesis using a verifier only in the inference time. We also show that typical error bounds based on the Rademacher complexity will be no larger than that of the original model when using a concurrent verifier in multi-class classification and structured prediction settings. Therefore, using a verifier in a learning phase will not hurt the generalizability of the model.

## [Recipe for a General, Powerful, Scalable Graph Transformer](#)

- Ladislav Rampářek · Mikhail Galkin · Vijay Prakash Dwivedi · Anh Tuan Luu · Guy Wolf · Dominique Beaini
- abstract@[open-review](#): We propose a recipe on how to build a general, powerful, scalable (GPS) graph Transformer with linear complexity and state-of-the-art results on a diverse set of benchmarks. Graph Transformers (GTs) have gained popularity in the field of graph representation learning with a variety of recent publications but they lack a common foundation about what constitutes a good positional or structural encoding, and what differentiates them. In this paper, we summarize the different types of encodings with a clearer definition and categorize them as being \$textit{local}\$, \$textit{global}\$ or \$textit{relative}\$. Further, GTs remain constrained to small graphs with few hundred nodes, and we propose the first architecture with a complexity linear to the number of nodes and edges \$O(N+E)\$ by decoupling the local real-edge aggregation from the fully-connected Transformer. We argue that this decoupling does not negatively affect the expressivity, with our architecture being a universal function approximator for graphs. Our GPS recipe consists of choosing 3 main ingredients: (i) positional/structural encoding, (ii) local message-passing mechanism, and (iii) global attention mechanism. We build and open-source a modular framework that supports multiple types of encodings and that provides efficiency and scalability both in small and large graphs. We test our architecture on 11 benchmarks and show very competitive results on all of them, showcasing the empirical benefits gained by the modularity and the combination of different strategies.

## [Hamiltonian Latent Operators for content and motion disentanglement in image sequences](#)

- Asif Khan · Amos Storkey
- abstract@[open-review](#): We introduce \textit{Halo} -- a deep generative model utilising HAmiltonian Latent Operators to disentangle content and motion information in image sequences reliably. The \textit{content} space captures summary statistics of a sequence, and \textit{motion} space under a dynamic process determines how information is expressed in any part of the sequence. By modelling the dynamics as a Hamiltonian motion, important desiderata are ensured: (1) the motion is reversible, (2) the symplectic, volume-preserving structure in phase space means paths are continuous and are not divergent in the space. Consequently, the nearness of sequence frames is realised by the nearness of their coordinates in the phase space, which proves valuable for long-term sequence generation. The sequence space is generally composed of different types of dynamical motions. To ensure long-term separability and perform controlled generation, we associate every motion with a unique Hamiltonian that acts in its respective subspace. We demonstrate the utility of our model by swapping the motion of a pair of sequences, controlled generation, and image rotations.

## [TA-GATES: An Encoding Scheme for Neural Network Architectures](#)

- Xuefei Ning · Zixuan Zhou · Junbo Zhao · Tianchen Zhao · Yiping Deng · Changcheng Tang · Shuang Liang · Huazhong Yang · Yu Wang
- abstract@[open-review](#): Neural architecture search tries to shift the manual design of neural network (NN) architectures to algorithmic design. In these cases, the NN architecture itself can be viewed as data and needs to be modeled. A better modeling could help explore novel architectures automatically and open the black box of automated architecture design. To this end, this work proposes a new encoding scheme for neural architectures, the Training-Analogous Graph-based ArchiTecture Encoding Scheme (TA-GATES). TA-GATES encodes an NN architecture in a way that is analogous to its training. Extensive experiments demonstrate that the flexibility and discriminative power of TA-GATES lead to better modeling of NN architectures. We expect our methodology of explicitly modeling the NN training process to benefit broader automated deep learning systems. The code is available at [https://github.com/walkerning/aw\\_nas](https://github.com/walkerning/aw_nas).

## [Learning Contrastive Embedding in Low-Dimensional Space](#)

- Shuo Chen · Chen Gong · Jun Li · Jian Yang · Gang Niu · Masashi Sugiyama
- abstract@[open-review](#): Contrastive learning (CL) pretrains feature embeddings to scatter instances in the feature space so that the training data can be well discriminated. Most existing CL techniques usually encourage learning such feature embeddings in the high-dimensional space to maximize the instance discrimination. However, this practice may result in the curse of dimensionality where the scattering instances are sparsely distributed in the high-dimensional feature space, making it difficult to capture the underlying similarity between pairwise instances. To this end, we propose a novel framework called contrastive learning with low-dimensional reconstruction (CLLR), which adopts a regularized projection layer to reduce the dimensionality of the feature embedding. In CLLR, we build the sparse low-rank regularizer to adaptively reconstruct a low-dimensional projection space while preserving the basic objective for instance discrimination, and thus successfully learning contrastive embeddings that alleviate the curse of dimensionality.

Theoretically, we prove a tighter error bound for CLLR; empirically, the superiority of CLLR is demonstrated across multiple domains, i.e., image classification, sentence representation, and reinforcement learning. Both theoretical and experimental results emphasize the significance of learning low-dimensional contrastive embeddings.

## [The trade-offs of model size in large recommendation models : A 10000 \\$times\\$ compressed criteo-tb DLRM model \(100 GB parameters to mere 10MB\)](#)

- Aditya Desai Â· Anshumali Shrivastava
- abstract@[open-review](#): Embedding tables dominate industrial-scale recommendation model sizes, using up to terabytes of memory. A popular and publicly available MLPerf benchmark on recommendation data is a Deep Learning Recommendation Model (DLRM) trained on a terabyte of click-through data. It contains 100GB of embedding memory (25+Billion parameters). DLRMs, due to their sheer size and the associated volume of data, face difficulty in training, deploying for inference, and memory bottlenecks due to large embedding tables. This paper analyzes and extensively evaluates a generic parameter sharing setup (PSS) for compressing DLRM models. We show theoretical upper bounds on the learnable memory requirements for achieving \$(1 \backslash pm \backslash epsilon)\$ approximations to the embedding table. Our bounds indicate exponentially fewer parameters suffice for good accuracy. To this end, we demonstrate a PSS DLRM reaching 10000\$times\$ compression on criteo-tb without losing quality. Such a compression, however, comes with a caveat. It requires 4.5 \$times\$ more iterations to reach the same saturation quality. The paper argues that this tradeoff needs more investigations as it might be significantly favorable. Leveraging the small size of the compressed model, we show a 4.3\$times\$ improvement in training latency leading to similar overall training times. Thus, in the tradeoff between system advantage of a small DLRM model vs. slower convergence, we show that scales are tipped towards having a smaller DLRM model, leading to faster inference, easier deployment, and similar training times.

## [Most Activation Functions Can Win the Lottery Without Excessive Depth](#)

- Rebekka Burkholz
- abstract@[open-review](#): The strong lottery ticket hypothesis has highlighted the potential for training deep neural networks by pruning, which has inspired interesting practical and theoretical insights into how neural networks can represent functions. For networks with ReLU activation functions, it has been proven that a target network with depth L can be approximated by the subnetwork of a randomly initialized neural network that has double the target's depth 2L and is wider by a logarithmic factor. We show that a depth L+1 is sufficient. This result indicates that we can expect to find lottery tickets at realistic, commonly used depths while only requiring logarithmic overparametrization. Our novel construction approach applies to a large class of activation functions and is not limited to ReLUs.

## [Generic bounds on the approximation error for physics-informed \(and\) operator learning](#)

- Tim De Ryck Â· Siddhartha Mishra
- abstract@[open-review](#): We propose a very general framework for deriving rigorous bounds on the approximation error for physics-informed neural networks (PINNs) and operator learning architectures such as DeepONets and FNOs as well as for physics-informed operator learning. These bounds guarantee that PINNs and (physics-informed) DeepONets or FNOs will efficiently approximate the underlying solution or solution-operator of generic partial differential equations (PDEs). Our framework utilizes existing neural network approximation results to obtain bounds on more-involved learning architectures for PDEs. We illustrate the general framework by deriving the first rigorous bounds on the approximation error of physics-informed operator learning and by showing that PINNs (and physics-informed DeepONets and FNOs) mitigate the curse of dimensionality in approximating nonlinear parabolic PDEs.

## [Active Learning for Multiple Target Models](#)

- Ying-Peng Tang Â· Sheng-Jun Huang
- abstract@[open-review](#): We describe and explore a novel setting of active learning (AL), where there are multiple target models to be learned simultaneously. In many real applications, the machine learning system is required to be deployed on diverse devices with varying computational resources (e.g., workstation, mobile phone, edge devices, etc.), which leads to the demand of training multiple target models on the same labeled dataset. However, it is generally believed that AL is model-dependent and untransferable, i.e., the data queried by one model may be less effective for training another model. This phenomenon naturally raises a question "Does there exist an AL method that is effective for multiple target models?" In this paper, we answer this question by theoretically analyzing the label complexity of active and passive learning under the setting with multiple target models, and conclude that AL does have potential to achieve better label complexity under this novel setting. Based on this insight, we further propose an agnostic AL sampling strategy to select the examples located in the joint disagreement regions of different target models. The experimental results on the OCR benchmarks show that the proposed method can significantly surpass the traditional active and passive learning methods under this challenging setting.

## [GLIPv2: Unifying Localization and Vision-Language Understanding](#)

- Haotian Zhang Â· Pengchuan Zhang Â· Xiaowei Hu Â· Yen-Chun Chen Â· Liunian Li Â· Xiyang Dai Â· Lijuan Wang Â· Lu Yuan Â· Jenq-Neng Hwang Â· Jianfeng Gao
- abstract@[open-review](#): We present GLIPv2, a grounded VL understanding model, that serves both localization tasks (e.g., object detection, instance segmentation) and Vision-Language (VL) understanding tasks (e.g., VQA, image captioning). GLIPv2 elegantly unifies localization pre-training and Vision-Language Pre-training (VLP) with three pre-training tasks: phrase grounding as a VL reformulation of the detection task, region-word contrastive learning as a novel region-word level contrastive learning task, and the masked language modeling. This unification not only simplifies the previous multi-stage VLP procedure but also achieves mutual benefits between localization and understanding tasks. Experimental results show that a single GLIPv2 model (all model weights are shared) achieves near SoTA performance on various localization and understanding tasks. The model also shows (1) strong zero-shot and few-shot adaption performance on open-vocabulary object detection tasks and (2) superior grounding capability on VL understanding tasks.

## [Neural Surface Reconstruction of Dynamic Scenes with Monocular RGB-D Camera](#)

- Hongrui Cai Â· Wanquan Feng Â· Xuetao Feng Â· Yan Wang Â· Juyong Zhang
- abstract@[open-review](#): We propose Neural-DynamicReconstruction (NDR), a template-free method to recover high-fidelity geometry and motions of a dynamic scene from a monocular RGB-D camera. In NDR, we adopt the neural implicit function for surface representation and rendering such that the captured color and depth can be fully utilized to jointly optimize the surface and deformations. To represent and constrain the non-rigid deformations, we propose a novel neural invertible deforming network such that the cycle consistency between arbitrary two frames is automatically satisfied. Considering that the surface topology of dynamic scene might change over time, we employ a topology-aware strategy to construct the topology-varient correspondence for the fused frames. NDR also further refines the camera poses in a global optimization manner. Experiments on public datasets and our collected dataset demonstrate that NDR outperforms existing monocular dynamic reconstruction methods.

## [Perceptual Attacks of No-Reference Image Quality Models with Human-in-the-Loop](#)

- Weixia Zhang Â· Dingquan Li Â· Xiongkuo Min Â· Guangtao Zhai Â· Guodong Guo Â· Xiaokang Yang Â· Kede Ma

- abstract@[open-review](#): No-reference image quality assessment (NR-IQA) aims to quantify how humans perceive visual distortions of digital images without access to their undistorted references. NR-IQA models are extensively studied in computational vision, and are widely used for performance evaluation and perceptual optimization of man-made vision systems. Here we make one of the first attempts to examine the perceptual robustness of NR-IQA models. Under a Lagrangian formulation, we identify insightful connections of the proposed perceptual attack to previous beautiful ideas in computer vision and machine learning. We test one knowledge-driven and three data-driven NR-IQA methods under four full-reference IQA models (as approximations to human perception of just-noticeable differences). Through carefully designed psychophysical experiments, we find that all four NR-IQA models are vulnerable to the proposed perceptual attack. More interestingly, we observe that the generated counterexamples are not transferable, manifesting themselves as distinct design flows of respective NR-IQA methods.

## [GT-GAN: General Purpose Time Series Synthesis with Generative Adversarial Networks](#)

- Jinsung Jeon · JEONGHAK KIM · Haryong Song · Seunghyeon Cho · Noseong Park
- abstract@[open-review](#): Time series synthesis is an important research topic in the field of deep learning, which can be used for data augmentation. Time series data types can be broadly classified into regular or irregular. However, there are no existing generative models that show good performance for both types without any model changes. Therefore, we present a general purpose model capable of synthesizing regular and irregular time series data. To our knowledge, we are the first designing a general purpose time series synthesis model, which is one of the most challenging settings for time series synthesis. To this end, we design a generative adversarial network-based method, where many related techniques are carefully integrated into a single framework, ranging from neural ordinary/controlled differential equations to continuous time-flow processes. Our method outperforms all existing methods.

## [Variational inference via Wasserstein gradient flows](#)

- Marc Lambert · Sinho Chewi · Francis Bach · Silvère Bonnabel · Philippe Rigollet
- abstract@[open-review](#): Along with Markov chain Monte Carlo (MCMC) methods, variational inference (VI) has emerged as a central computational approach to large-scale Bayesian inference. Rather than sampling from the true posterior  $\pi$ , VI aims at producing a simple but effective approximation  $\hat{\pi}$  to  $\pi$  for which summary statistics are easy to compute. However, unlike the well-studied MCMC methodology, VI is still poorly understood and dominated by heuristics. In this work, we propose principled methods for VI, in which  $\hat{\pi}$  is taken to be a Gaussian or a mixture of Gaussians, which rest upon the theory of gradient flows on the Bures--Wasserstein space of Gaussian measures. Akin to MCMC, it comes with strong theoretical guarantees when  $\pi$  is log-concave.

## [Discrete-Convex-Analysis-Based Framework for Warm-Starting Algorithms with Predictions](#)

- Shinsaku Sakaue · Taihei Oki
- abstract@[open-review](#): Augmenting algorithms with learned predictions is a promising approach for going beyond worst-case bounds. Dinitz, Im, Lavastida, Moseley, and Vassilvitskii~(2021) have demonstrated that a warm start with learned dual solutions can improve the time complexity of the Hungarian method for weighted perfect bipartite matching. We extend and improve their framework in a principled manner via discrete convex analysis (DCA), a discrete analog of convex analysis. We show the usefulness of our DCA-based framework by applying it to weighted perfect bipartite matching, weighted matroid intersection, and discrete energy minimization for computer vision. Our DCA-based framework yields time complexity bounds that depend on the  $\ell_\infty$ -distance from a predicted solution to an optimal solution, which has two advantages relative to the previous  $\ell_1$ -distance-dependent bounds: time complexity bounds are smaller, and learning of predictions is more sample efficient. We also discuss whether to learn primal or dual solutions from the DCA perspective.

## [Fast Instrument Learning with Faster Rates](#)

- Ziyu Wang · Yuhao Zhou · Jun Zhu
- abstract@[open-review](#): We investigate nonlinear instrumental variable (IV) regression given high-dimensional instruments. We propose a simple algorithm which combines kernelized IV methods and an arbitrary, adaptive regression algorithm, accessed as a black box. Our algorithm enjoys faster-rate convergence and adapts to the dimensionality of informative latent features, while avoiding an expensive minimax optimization procedure, which has been necessary to establish similar guarantees. It further brings the benefit of flexible machine learning models to quasi-Bayesian uncertainty quantification, likelihood-based model selection, and model averaging. Simulation studies demonstrate the competitive performance of our method.

## [Double Check Your State Before Trusting It: Confidence-Aware Bidirectional Offline Model-Based Imagination](#)

- Jiafei Lyu · Xiu Li · Zongqing Lu
- abstract@[open-review](#): The learned policy of model-free offline reinforcement learning (RL) methods is often constrained to stay within the support of datasets to avoid possible dangerous out-of-distribution actions or states, making it challenging to handle out-of-support region. Model-based RL methods offer a richer dataset and benefit generalization by generating imaginary trajectories with either trained forward or reverse dynamics model. However, the imagined transitions may be inaccurate, thus downgrading the performance of the underlying offline RL method. In this paper, we propose to augment the offline dataset by using trained bidirectional dynamics models and rollout policies with double check. We introduce conservatism by trusting samples that the forward model and backward model agree on. Our method, confidence-aware bidirectional offline model-based imagination, generates reliable samples and can be combined with any model-free offline RL method. Experimental results on the D4RL benchmarks demonstrate that our method significantly boosts the performance of existing model-free offline RL algorithms and achieves competitive or better scores against baseline methods.

## [OGC: Unsupervised 3D Object Segmentation from Rigid Dynamics of Point Clouds](#)

- Ziyang Song · Bo Yang
- abstract@[open-review](#): In this paper, we study the problem of 3D object segmentation from raw point clouds. Unlike all existing methods which usually require a large amount of human annotations for full supervision, we propose the first unsupervised method, called OGC, to simultaneously identify multiple 3D objects in a single forward pass, without needing any type of human annotations. The key to our approach is to fully leverage the dynamic motion patterns over sequential point clouds as the supervision signals to automatically discover rigid objects. Our method consists of three major components, 1) the object segmentation network to directly estimate multi-object masks from a single point cloud frame, 2) the auxiliary self-supervised scene flow estimator, and 3) our core object geometry consistency component. By carefully designing a series of loss functions, we effectively take into account the multi-object rigid consistency and the object shape invariance in both temporal and spatial scales. This allows our method to truly discover the object geometry even in the absence of annotations. We extensively evaluate our method on four datasets, demonstrating the superior performance for object part instance segmentation and general object segmentation in both indoor and the challenging outdoor scenarios.

## [Learning Distributions Generated by Single-Layer ReLU Networks in the Presence of Arbitrary Outliers](#)

- Saikiran Bulusu · Geethu Joseph · M. Cenk Gursoy · Pramod Varshney
- abstract@[open-review](#): We consider a set of data samples such that a fraction of the samples are arbitrary outliers, and the rest are the output samples of a single-layer neural network with rectified linear unit (ReLU) activation. Our goal is to estimate the parameters (weight matrix and bias vector) of the

neural network, assuming the bias vector to be non-negative. We estimate the network parameters using the gradient descent algorithm combined with either the median- or trimmed mean-based filters to mitigate the effect of the arbitrary outliers. We then prove that  $\tilde{O}(\frac{1}{p^2} + \frac{1}{\epsilon^{2p}})$  samples and  $\tilde{O}(\frac{d^2}{p^2} + \frac{d^2}{\epsilon^{2p}})$  time are sufficient for our algorithm to estimate the neural network parameters within an error of  $\epsilon$  when the outlier probability is  $1-p$ , where  $p/2$

## [Alleviating the Sampling Bias of Few Shot Data by Removing Projection to the Centroid](#)

- Jing Xu · Xu Luo · Xinglin Pan · Yanan Li · Wenjie Pei · Zenglin Xu
- abstract@[open-review](#): Few-shot learning (FSL) aims to achieve good generalization without sufficient annotations in the novel classes. Despite the successes of a number of few-shot learning methods motivated from various perspectives, the sensitivity to the limited amount and the discriminative power of support data is not well understood, which is also called the sampling bias problem. This paper reveals one such phenomenon ---- the classification boundary is very sensitive to the position of support samples if they are in the vicinity of the data centroid, which we call the task centroid expressing the data centroids for a given task, degenerated and unstable results are usually observed. To reduce this sampling bias, motivated by the effect of the task centroid, we propose a simple feature transformation, named Task Centroid Projection Removing(TCPR). TCPR aims to remove the component of features along the direction of approximated task centroid which is estimated through similar examples from the base dataset. This effectively prevents features from being too close to the task centroid. Extensive experiments over ten datasets from different domains show that TCPR can reliably improve classification accuracy across various feature extractors, training algorithms, and datasets. The code can be found in the Supplementary.

## [Don't Roll the Dice, Ask Twice: The Two-Query Distortion of Matching Problems and Beyond](#)

- Georgios Amanatidis · Georgios Birmpas · Aris Filos-Ratsikas · Alexandros Voudouris
- abstract@[open-review](#): In most social choice settings, the participating agents express their preferences over the different alternatives in the form of linear orderings. While this clearly simplifies preference elicitation, it inevitably leads to poor performance with respect to optimizing a cardinal objective, such as the social welfare, since the values of the agents remain virtually unknown. This loss in performance because of lack of information is measured by distortion. A recent array of works put forward the agenda of designing mechanisms that learn the values of the agents for a small number of alternatives via queries, and use this limited extra information to make better-informed decisions, thus improving distortion. Following this agenda, in this work we focus on a class of combinatorial problems that includes most well-known matching problems and several of their generalizations, such as One-Sided Matching, Two-Sided Matching, General Graph Matching, and k-Constrained Resource Allocation. We design two-query mechanisms that achieve the best-possible worst-case distortion in terms of social welfare, and outperform the best-possible expected distortion achieved by randomized ordinal mechanisms.

## [Generalization Bounds for Estimating Causal Effects of Continuous Treatments](#)

- Xin Wang · Shengfei Lyu · Xingyu Wu · Tianhao Wu · Huanhuan Chen
- abstract@[open-review](#): We focus on estimating causal effects of continuous treatments (e.g., dosage in medicine), also known as dose-response function. Existing methods in causal inference for continuous treatments using neural networks are effective and to some extent reduce selection bias, which is introduced by non-randomized treatments among individuals and might lead to covariate imbalance and thus unreliable inference. To theoretically support the alleviation of selection bias in the setting of continuous treatments, we exploit the re-weighting schema and the Integral Probability Metric (IPM) distance to derive an upper bound on the counterfactual error when estimating the average dose-response function (ADRF). The IPM distance builds a bridge between a source (factual) and an infinite number of target (counterfactual) domains. We provide a discretized approximation of the IPM distance with a theoretical guarantee in the practical implementation. Based on the theoretical analysis, we propose a novel algorithm, called Average Dose-response estiMatIon via re-weighTing schema (ADMIT), which simultaneously learns a re-weighting network, which aims to alleviate the selection bias, and an inference network, which makes factual and counterfactual estimations. Besides, the effectiveness of ADMIT is empirically indicated in both synthetic and semi-synthetic experiments by outperforming the existing benchmarks.

## [MaskTune: Mitigating Spurious Correlations by Forcing to Explore](#)

- Saeid Asgari · Aliashghar Khani · Fereshte Khani · Ali Gholami · Linh Tran · Ali Mahdavi-Amiri · Ghassan Hamarneh
- abstract@[open-review](#): A fundamental challenge of over-parameterized deep learning models is learning meaningful data representations that yield good performance on a downstream task without over-fitting spurious input features. This work proposes MaskTune, a masking strategy that prevents over-reliance on spurious (or a limited number of) features. MaskTune forces the trained model to explore new features during a single epoch finetuning by masking previously discovered features. MaskTune, unlike earlier approaches for mitigating shortcut learning, does not require any supervision, such as annotating spurious features or labels for subgroup samples in a dataset. Our empirical results on biased MNIST, CelebA, Waterbirds, and ImagenNet-9L datasets show that MaskTune is effective on tasks that often suffer from the existence of spurious correlations. Finally, we show that our method outperforms or achieves similar performance to the competing methods when applied to the selective classification task.

## [Neuron with Steady Response Leads to Better Generalization](#)

- Qiang Fu · Lun Du · Haitao Mao · Xu Chen · Wei Fang · Shi Han · Dongmei Zhang
- abstract@[open-review](#): Regularization can mitigate the generalization gap between training and inference by introducing inductive bias. Existing works have already proposed various inductive biases from diverse perspectives. However, none of them explores inductive bias from the perspective of class-dependent response distribution of individual neurons. In this paper, we conduct a substantial analysis of the characteristics of such distribution. Based on the analysis results, we articulate the Neuron Steadiness Hypothesis: the neuron with similar responses to instances of the same class leads to better generalization. Accordingly, we propose a new regularization method called Neuron Steadiness Regularization (NSR) to reduce neuron intra-class response variance. Based on Complexity Measure, we theoretically guarantee the effectiveness of NSR for improving generalization. We conduct extensive experiments on Multilayer Perceptron, Convolutional Neural Network, and Graph Neural Network with popular benchmark datasets of diverse domains, which show that our Neuron Steadiness Regularization consistently outperforms the vanilla version of models with significant gain and low additional computational overhead.

## [Improved Feature Distillation via Projector Ensemble](#)

- Yudong Chen · Sen Wang · Jiajun Liu · Xuwei Xu · Frank de Hoog · Zi Huang
- abstract@[open-review](#): Knowledge Distillation has been widely used to improve the performance of the lightweight network (student) by introducing the large network (teacher) to guide training. Among the existing methods, feature matching-based distillation has shown superior performance by minimizing the discrepancy between student and teacher features. Due to the dimension mismatch between student and teacher features, feature distillation methods usually impose a projector on the student or teacher networks to map features into a common space during training. Previous feature distillation methods mainly focus on the design of loss functions and the selection of the distilled layers, while the effect of the feature projector between the student and teacher remains under-explored. To better understand the impact of projectors in distillation, we conduct comprehensive experiments in this paper and observe that the student network benefits from a projector even if the feature dimensions of the student and teacher are the same. One plausible reason is that the projector is optimised towards a "global alignment" that cannot be achieved by just optimising independent feature pairs. Motivated by this, we

propose an ensemble of projectors to further improve the distillation performance. Empirical results on a series of teacher-student pairs illustrate the effectiveness of the proposed method.

## [Rotation-Equivariant Conditional Spherical Neural Fields for Learning a Natural Illumination Prior](#)

- James Gardner · Bernhard Egger · William Smith
- abstract@[open-review](#): Inverse rendering is an ill-posed problem. Previous work has sought to resolve this by focussing on priors for object or scene shape or appearance. In this work, we instead focus on a prior for natural illuminations. Current methods rely on spherical harmonic lighting or other generic representations and, at best, a simplistic prior on the parameters. We propose a conditional neural field representation based on a variational auto-decoder with a SIREN network and, extending Vector Neurons, build equivariance directly into the network. Using this we develop a rotation-equivariant, high dynamic range (HDR) neural illumination model that is compact and able to express complex, high-frequency features of natural environment maps. Training our model on a curated dataset of 1.6K HDR environment maps of natural scenes, we compare it against traditional representations, demonstrate its applicability for an inverse rendering task and show environment map completion from partial observations. A PyTorch implementation, our dataset and trained models will be made available.

## [Why neural networks find simple solutions: The many regularizers of geometric complexity](#)

- Benoit Dherin · Michael Munn · Mihaela Rosca · David Barrett
- abstract@[open-review](#): In many contexts, simpler models are preferable to more complex models and the control of this model complexity is the goal for many methods in machine learning such as regularization, hyperparameter tuning and architecture design. In deep learning, it has been difficult to understand the underlying mechanisms of complexity control, since many traditional measures are not naturally suitable for deep neural networks. Here we develop the notion of geometric complexity, which is a measure of the variability of the model function, computed using a discrete Dirichlet energy. Using a combination of theoretical arguments and empirical results, we show that many common training heuristics such as parameter norm regularization, spectral norm regularization, flatness regularization, implicit gradient regularization, noise regularization and the choice of parameter initialization all act to control geometric complexity, providing a unifying framework in which to characterize the behavior of deep learning models.

## [Will Bilevel Optimizers Benefit from Loops](#)

- Kaiyi Ji · Mingrui Liu · Yingbin Liang · Lei Ying
- abstract@[open-review](#): Bilevel optimization has arisen as a powerful tool for solving a variety of machine learning problems. Two current popular bilevel optimizers AID-BiO and ITD-BiO naturally involve solving one or two sub-problems, and consequently, whether we solve these problems with loops (that take many iterations) or without loops (that take only a few iterations) can significantly affect the overall computational efficiency. Existing studies in the literature cover only some of those implementation choices, and the complexity bounds available are not refined enough to enable rigorous comparison among different implementations. In this paper, we first establish unified convergence analysis for both AID-BiO and ITD-BiO that are applicable to all implementation choices of loops. We then specialize our results to characterize the computational complexity for all implementations, which enable an explicit comparison among them. Our result indicates that for AID-BiO, the loop for estimating the optimal point of the inner function is beneficial for overall efficiency, although it causes higher complexity for each update step, and the loop for approximating the outer-level Hessian-inverse-vector product reduces the gradient complexity. For ITD-BiO, the two loops always coexist, and our convergence upper and lower bounds show that such loops are necessary to guarantee a vanishing convergence error, whereas the no-loop scheme suffers from an unavoidable non-vanishing convergence error. Our numerical experiments further corroborate our theoretical results.

## [Sequencer: Deep LSTM for Image Classification](#)

- Yuki Tatsunami · Masato Taki
- abstract@[open-review](#): In recent computer vision research, the advent of the Vision Transformer (ViT) has rapidly revolutionized various architectural design efforts: ViT achieved state-of-the-art image classification performance using self-attention found in natural language processing, and MLP-Mixer achieved competitive performance using simple multi-layer perceptrons. In contrast, several studies have also suggested that carefully redesigned convolutional neural networks (CNNs) can achieve advanced performance comparable to ViT without resorting to these new ideas. Against this background, there is growing interest in what inductive bias is suitable for computer vision. Here we propose Sequencer, a novel and competitive architecture alternative to ViT that provides a new perspective on these issues. Unlike ViTs, Sequencer models long-range dependencies using LSTMs rather than self-attention layers. We also propose a two-dimensional version of Sequencer module, where an LSTM is decomposed into vertical and horizontal LSTMs to enhance performance. Despite its simplicity, several experiments demonstrate that Sequencer performs impressively well: Sequencer2D-L, with 54M parameters, realizes 84.6% top-1 accuracy on only ImageNet-1K. Not only that, we show that it has good transferability and the robust resolution adaptability on double resolution-band.

## [S-PIFu: Integrating Parametric Human Models with PIFu for Single-view Clothed Human Reconstruction](#)

- Kennard Chan · Guosheng Lin · Haiyu Zhao · Weisi Lin
- abstract@[open-review](#): We present three novel strategies to incorporate a parametric body model into a pixel-aligned implicit model for single-view clothed human reconstruction. Firstly, we introduce ray-based sampling, a novel technique that transforms a parametric model into a set of highly informative, pixel-aligned 2D feature maps. Next, we propose a new type of feature based on blendweights. Blendweight-based labels serve as soft human parsing labels and can help to significantly improve the structural fidelity of reconstructed meshes. Finally, we show how we can extract and capitalize on body part orientation information from a parametric model to further improve reconstruction quality. Together, these three techniques form our S-PIFu framework, which significantly outperforms state-of-the-arts methods in all metrics.

## [DISCO: Adversarial Defense with Local Implicit Functions](#)

- Chih-Hui Ho · Nuno Vasconcelos
- abstract@[open-review](#): The problem of adversarial defenses for image classification, where the goal is to robustify a classifier against adversarial examples, is considered. Inspired by the hypothesis that these examples lie beyond the natural image manifold, a novel aDversarIal defenSe with local impliCIt functiOns (DISCO) is proposed to remove adversarial perturbations by localized manifold projections. DISCO consumes an adversarial image and a query pixel location and outputs a clean RGB value at the location. It is implemented with an encoder and a local implicit module, where the former produces per-pixel deep features and the latter uses the features in the neighborhood of query pixel for predicting the clean RGB value. Extensive experiments demonstrate that both DISCO and its cascade version outperform prior defenses, regardless of whether the defense is known to the attacker. DISCO is also shown to be data and parameter efficient and to mount defenses that transfers across datasets, classifiers and attacks.

## [Misspecified Phase Retrieval with Generative Priors](#)

- Zhaoqiang Liu · Xinshao Wang · Jiulong Liu
- abstract@[open-review](#): In this paper, we study phase retrieval under model misspecification and generative priors. In particular, we aim to estimate an  $n$ -dimensional signal  $\mathbf{x}$  from  $m$  i.i.d. realizations of the single index model  $y = f(\mathbf{a})^T \mathbf{x}$ , where  $f$  is an unknown

and possibly random nonlinear link function and  $\mathbf{a} \in \mathbb{R}^n$  is a standard Gaussian vector. We make the assumption  $\mathrm{Cov}[y, (\mathbf{a})^T \mathbf{a}] \neq 0$ , which corresponds to the misspecified phase retrieval problem. In addition, the underlying signal  $\mathbf{x}$  is assumed to lie in the range of an  $L$ -Lipschitz continuous generative model with bounded  $k$ -dimensional inputs. We propose a two-step approach, for which the first step plays the role of spectral initialization and the second step refines the estimated vector produced by the first step iteratively. We show that both steps enjoy a statistical rate of order  $\sqrt{(k \log L) \cdot (\log m)/m}$  under suitable conditions. Experiments on image datasets are performed to demonstrate that our approach performs on par with or even significantly outperforms several competing methods.

## [Target alignment in truncated kernel ridge regression](#)

- Arash Amini · Richard Baumgartner · Dai Feng
- abstract@[open-review](#): Kernel ridge regression (KRR) has recently attracted renewed interest due to its potential for explaining the transient effects, such as double descent, that emerge during neural network training. In this work, we study how the alignment between the target function and the kernel affects the performance of the KRR. We focus on the truncated KRR (TKRR) which utilizes an additional parameter that controls the spectral truncation of the kernel matrix. We show that for polynomial alignment, there is an over-aligned regime, in which TKRR can achieve a faster rate than what is achievable by full KRR. The rate of TKRR can improve all the way to the parametric rate, while that of full KRR is capped at a sub-optimal value. This shows that target alignment can be better leveraged by utilizing spectral truncation in kernel methods. We also consider the bandlimited alignment setting and show that the regularization surface of TKRR can exhibit transient effects including multiple descent and non-monotonic behavior. Our results show that there is a strong and quantifiable relation between the shape of the alignment spectrum and the generalization performance of kernel methods, both in terms of rates and in finite samples.

## [Uncertainty Estimation Using Riemannian Model Dynamics](#)

- Guy Tennenholz · Shie Mannor
- abstract@[open-review](#): Model-based offline reinforcement learning approaches generally rely on bounds of model error. Estimating these bounds is usually achieved through uncertainty estimation methods. In this work, we combine parametric and nonparametric methods for uncertainty estimation through a novel latent space based metric. In particular, we build upon recent advances in Riemannian geometry of generative models to construct a pullback metric of an encoder-decoder based forward model. Our proposed metric measures both the quality of out-of-distribution samples as well as the discrepancy of examples in the data. We leverage our combined method for uncertainty estimation in a pessimistic model-based framework, showing a significant improvement upon contemporary model-based offline approaches on continuous control and autonomous driving benchmarks.

## [High-Order Pooling for Graph Neural Networks with Tensor Decomposition](#)

- Chenqing Hua · Guillaume Rabusseau · Jian Tang
- abstract@[open-review](#): Graph Neural Networks (GNNs) are attracting growing attention due to their effectiveness and flexibility in modeling a variety of graph-structured data. Existing GNN architectures usually adopt simple pooling operations (e.g., sum, average, max) when aggregating messages from a local neighborhood for updating node representation or pooling node representations from the entire graph to compute the graph representation. Though simple and effective, these linear operations ignore modeling the high-order non-linear interactions among nodes, which limits their expressivity. In this paper, we propose the Tensorized Graph Neural Network (tGNN), a highly expressive GNN architecture modeling high-order non-linear node interactions based on symmetric tensor decomposition. tGNN leverages the symmetric CP decomposition to efficiently parameterize permutation-invariant multilinear maps for modeling node interactions. Theoretical and empirical analysis on both node and graph classification tasks show the superiority of our proposed tGNN over competitive baselines. In particular, tGNN achieves state-of-the-art results on two OGB node classification datasets and one OGB graph classification dataset.

## [BMU-MoCo: Bidirectional Momentum Update for Continual Video-Language Modeling](#)

- Yizhao Gao · Nanyi Fei · Haoyu Lu · Zhiwu Lu · Hao Jiang · Yijie Li · Zhao Cao
- abstract@[open-review](#): Video-language models suffer from forgetting old/learned knowledge when trained with streaming data. In this work, we thus propose a continual video-language modeling (CVLM) setting, where models are supposed to be sequentially trained on five widely-used video-text datasets with different data distributions. Although most of existing continual learning methods have achieved great success by exploiting extra information (e.g., memory data of past tasks) or dynamically extended networks, they cause enormous resource consumption when transferred to our CVLM setting. To overcome the challenges (i.e., catastrophic forgetting and heavy resource consumption) in CVLM, we propose a novel cross-modal MoCo-based model with bidirectional momentum update (BMU), termed BMU-MoCo. Concretely, our BMU-MoCo has two core designs: (1) Different from the conventional MoCo, we apply the momentum update to not only momentum encoders but also encoders (i.e., bidirectional) at each training step, which enables the model to review the learned knowledge retained in the momentum encoders. (2) To further enhance our BMU-MoCo by utilizing earlier knowledge, we additionally maintain a pair of global momentum encoders (only initialized at the very beginning) with the same BMU strategy. Extensive results show that our BMU-MoCo remarkably outperforms recent competitors w.r.t. video-text retrieval performance and forgetting rate, even without using any extra data or dynamic networks.

## [Bivariate Causal Discovery for Categorical Data via Classification with Optimal Label Permutation](#)

- Yang Ni
- abstract@[open-review](#): Causal discovery for quantitative data has been extensively studied but less is known for categorical data. We propose a novel causal model for categorical data based on a new classification model, termed classification with optimal label permutation (COLP). By design, COLP is a parsimonious classifier, which gives rise to a provably identifiable causal model. A simple learning algorithm via comparing likelihood functions of causal and anti-causal models suffices to learn the causal direction. Through experiments with synthetic and real data, we demonstrate the favorable performance of the proposed COLP-based causal model compared to state-of-the-art methods. We also make available an accompanying R package COLP, which contains the proposed causal discovery algorithm and a benchmark dataset of categorical cause-effect pairs.

## [Adversarial Reprogramming Revisited](#)

- Matthias Englert · Ranko Lazic
- abstract@[open-review](#): Adversarial reprogramming, introduced by Elsayed, Goodfellow, and Sohl-Dickstein, seeks to repurpose a neural network to perform a different task, by manipulating its input without modifying its weights. We prove that two-layer ReLU neural networks with random weights can be adversarially reprogrammed to achieve arbitrarily high accuracy on Bernoulli data models over hypercube vertices, provided the network width is no greater than its input dimension. We also substantially strengthen a recent result of Phuong and Lampert on directional convergence of gradient flow, and obtain as a corollary that training two-layer ReLU neural networks on orthogonally separable datasets can cause their adversarial reprogramming to fail. We support these theoretical results by experiments that demonstrate that, as long as batch normalisation layers are suitably initialised, even untrained networks with random weights are susceptible to adversarial reprogramming. This is in contrast to observations in several recent works that suggested that adversarial reprogramming is not possible for untrained networks to any degree of reliability.

## [Efficient and Effective Augmentation Strategy for Adversarial Training](#)

- Sravanti Addepalli · Samyak Jain · Venkatesh Babu R
- abstract@[open-review](#): The sample complexity of Adversarial training is known to be significantly higher than standard ERM based training. Although complex augmentation techniques have led to large gains in standard training, they have not been successful with Adversarial Training. In this work, we propose Diverse Augmentation based Joint Adversarial Training (DAJAT) that uses a combination of simple and complex augmentations with separate batch normalization layers to handle the conflicting goals of enhancing the diversity of the training dataset, while being close to the test distribution. We further introduce a Jensen-Shannon divergence loss to encourage the joint learning of the diverse augmentations, thereby allowing simple augmentations to guide the learning of complex ones. Lastly, to improve the computational efficiency of the proposed method, we propose and utilize a two-step defense, Ascending Constraint Adversarial Training (ACAT) that uses an increasing epsilon schedule and weight-space smoothing to prevent gradient masking. The proposed method achieves a better robustness-accuracy trade-off compared to existing methods on the RobustBench Leaderboard for CIFAR-10 and CIFAR-100 on ResNet-18 and WideResNet-34-10 architectures.

## [Making Sense of Dependence: Efficient Black-box Explanations Using Dependence Measure](#)

- Paul Novello · Thomas FEL · David Vigouroux
- abstract@[open-review](#): This paper presents a new efficient black-box attribution method based on Hilbert-Schmidt Independence Criterion (HSIC), a dependence measure based on Reproducing Kernel Hilbert Spaces (RKHS). HSIC measures the dependence between regions of an input image and the output of a model based on kernel embeddings of distributions. It thus provides explanations enriched by RKHS representation capabilities. HSIC can be estimated very efficiently, significantly reducing the computational cost compared to other black-box attribution methods. Our experiments show that HSIC is up to 8 times faster than the previous best black-box attribution methods while being as faithful. Indeed, we improve or match the state-of-the-art of both black-box and white-box attribution methods for several fidelity metrics on Imagenet with various recent model architectures. Importantly, we show that these advances can be transposed to efficiently and faithfully explain object detection models such as YOLOv4. Finally, we extend the traditional attribution methods by proposing a new kernel enabling an ANOVA-like orthogonal decomposition of importance scores based on HSIC, allowing us to evaluate not only the importance of each image patch but also the importance of their pairwise interactions. Our implementation is available at \url{https://github.com/paulnovello/HSIC-Attribution-Method}.

## [VCT: A Video Compression Transformer](#)

- Fabian Mentzer · George D Toderici · David Minnen · Sergi Caelles · Sung Jin Hwang · Mario Lucic · Eirikur Agustsson
- abstract@[open-review](#): We show how transformers can be used to vastly simplify neural video compression. Previous methods have been relying on an increasing number of architectural biases and priors, including motion prediction and warping operations, resulting in complex models. Instead, we independently map input frames to representations and use a transformer to model their dependencies, letting it predict the distribution of future representations given the past. The resulting video compression transformer outperforms previous methods on standard video compression data sets. Experiments on synthetic data show that our model learns to handle complex motion patterns such as panning, blurring and fading purely from data. Our approach is easy to implement, and we release code to facilitate future research.

## [Pragmatically Learning from Pedagogical Demonstrations in Multi-Goal Environments](#)

- Hugo Caselles-Dupré · Olivier Sigaud · Mohamed CHETOUANI
- abstract@[open-review](#): Learning from demonstration methods usually leverage close to optimal demonstrations to accelerate training. By contrast, when demonstrating a task, human teachers deviate from optimal demonstrations and pedagogically modify their behavior by giving demonstrations that best disambiguate the goal they want to demonstrate. Analogously, human learners excel at pragmatically inferring the intent of the teacher, facilitating communication between the two agents. These mechanisms are critical in the few demonstrations regime, where inferring the goal is more difficult. In this paper, we implement pedagogy and pragmatism mechanisms by leveraging a Bayesian model of goal inference from demonstrations. We highlight the benefits of this model in multi-goal teacher-learner setups with two artificial agents that learn with goal-conditioned Reinforcement Learning. We show that combining a pedagogical teacher and a pragmatic learner results in faster learning and reduced goal ambiguity over standard learning from demonstrations, especially in the few demonstrations regime.

## [Instance-based Learning for Knowledge Base Completion](#)

- Wanyun Cui · Xingran Chen
- abstract@[open-review](#): In this paper, we proposed a new method for knowledge base completion (KBC): instance-based learning (IBL). For example, to answer (Jill Biden, lived city, ?), instead of going directly to Washington D.C., our goal is to find Joe Biden, who has the same lived city as Jill Biden. Through prototype entities, IBL provides interpretability. We developed theories for modeling prototypes and combining IBL with translational models. Experiments on various tasks have confirmed the IBL model's effectiveness and interpretability. In addition, IBL shed light on the mechanism of rule-based KBC models. Previous research has generally agreed that rule-based methods provide rules with semantically related premise and hypothesis. We challenge this view. We begin by demonstrating that some logical rules represent \{\it instance-based equivalence\} (i.e. prototypes) rather than semantic relevance. These are denoted as \{\it IBL rules\}. Surprisingly, despite occupying only a small portion of the rule space, IBL rules outperform non-IBL rules in all four benchmarks. We use a variety of experiments to demonstrate that rule-based models work because they have the ability to represent instance-based equivalence via IBL rules. The findings provide new insights of how rule-based models work and how to interpret their rules.

## [VRL3: A Data-Driven Framework for Visual Deep Reinforcement Learning](#)

- Che Wang · Xufang Luo · Keith Ross · Dongsheng Li
- abstract@[open-review](#): We propose VRL3, a powerful data-driven framework with a minimalist design for solving highly challenging visual deep reinforcement learning (DRL) tasks. We analyze a number of major obstacles in taking a data-driven approach, and present a suite of design principles, novel findings, and critical insights about data-driven visual DRL. Our framework has three stages: in stage 1, we leverage non-RL datasets (e.g. ImageNet) to learn task-agnostic visual representations; in stage 2, we use offline RL data (e.g. a limited number of expert demonstrations) to convert the task-agnostic representations into more powerful task-specific representations; in stage 3, we fine-tune the agent with online RL. On a set of highly challenging hand manipulation tasks with sparse reward and realistic visual inputs, compared to the previous SOTA, VRL3 achieves an average of 780% better sample efficiency. And on the hardest task, VRL3 is 1220% more sample efficient and solves the task with only 10% of the computation. These highly significant results clearly demonstrate the great potential of data-driven deep reinforcement learning.

## [On the Convergence Theory for Hessian-Free Bilevel Algorithms](#)

- Daouda Sow · Kaiyi Ji · Yingbin Liang
- abstract@[open-review](#): Bilevel optimization has arisen as a powerful tool in modern machine learning. However, due to the nested structure of bilevel optimization, even gradient-based methods require second-order derivative approximations via Jacobian- or/and Hessian-vector computations, which can be costly and unscalable in practice. Recently, Hessian-free bilevel schemes have been proposed to resolve this issue, where the general idea is to use zeroth- or first-order methods to approximate the full hypergradient of the bilevel problem. However, we empirically observe that such approximation can lead to large variance and unstable training, but estimating only the response Jacobian matrix as a partial component of the hypergradient turns out to be extremely effective. To this end, we propose a new Hessian-free method, which adopts the zeroth-order-like method to approximate the response Jacobian matrix via taking difference between two optimization paths. Theoretically, we provide the convergence rate analysis for the proposed algorithms, where

our key challenge is to characterize the approximation and smoothness properties of the trajectory-dependent estimator, which can be of independent interest. This is the first known convergence rate result for this type of Hessian-free bilevel algorithms. Experimentally, we demonstrate that the proposed algorithms outperform baseline bilevel optimizers on various bilevel problems. Particularly, in our experiment on few-shot meta-learning with ResNet-12 network over the miniImageNet dataset, we show that our algorithm outperforms baseline meta-learning algorithms, while other baseline bilevel optimizers do not solve such meta-learning problems within a comparable time frame.

## [Towards Diverse and Faithful One-shot Adaption of Generative Adversarial Networks](#)

- Yabo Zhang · mingshuai Yao · Yuxiang Wei · Zhilong Ji · Jinfeng Bai · Wangmeng Zuo
- abstract@[open-review](#): One-shot generative domain adaption aims to transfer a pre-trained generator on one domain to a new domain using one reference image only. However, it remains very challenging for the adapted generator (i) to generate diverse images inherited from the pre-trained generator while (ii) faithfully acquiring the domain-specific attributes and styles of the reference image. In this paper, we present a novel one-shot generative domain adaption method, i.e., DiFa, for diverse generation and faithful adaptation. For global-level adaptation, we leverage the difference between the CLIP embedding of the reference image and the mean embedding of source images to constrain the target generator. For local-level adaptation, we introduce an attentive style loss which aligns each intermediate token of an adapted image with its corresponding token of the reference image. To facilitate diverse generation, selective cross-domain consistency is introduced to select and retain domain-sharing attributes in the editing latent  $\mathcal{W}$  space to inherit the diversity of the pre-trained generator. Extensive experiments show that our method outperforms the state-of-the-arts both quantitatively and qualitatively, especially for the cases of large domain gap. Moreover, our DiFa can easily be extended to zero-shot generative domain adaption with appealing results.

## [Pay attention to your loss : understanding misconceptions about Lipschitz neural networks](#)

- Louis Bøthune · Thibaut Boissin · Mathieu Serrurier · Franck Mamalet · Corentin Friedrich · Alberto Gonzalez Sanz
- abstract@[open-review](#): Lipschitz constrained networks have gathered considerable attention in deep learning community, with usages ranging from Wasserstein distance estimation to the training of certifiably robust classifiers. However they remain commonly considered as less accurate, and their properties in learning are still not fully understood. In this paper we clarify the matter: when it comes to classification 1-Lipschitz neural networks enjoy several advantages over their unconstrained counterpart. First, we show that these networks are as accurate as classical ones, and can fit arbitrarily difficult boundaries. Then, relying on a robustness metric which reflects operational needs we characterize the most robust classifier: the WGAN discriminator. Next, we show that 1-Lipschitz neural networks generalize well under milder assumptions. Finally, we show that hyper-parameters of the loss are crucial for controlling the accuracy-robustness trade-off. We conclude that they exhibit appealing properties to pave the way toward provably accurate, and provably robust neural networks.

## [GAR: Generalized Autoregression for Multi-Fidelity Fusion](#)

- Yuxin Wang · Zheng Xing · WEI XING
- abstract@[open-review](#): In many scientific research and engineering applications, where repeated simulations of complex systems are conducted, a surrogate is commonly adopted to quickly estimate the whole system. To reduce the expensive cost of generating training examples, it has become a promising approach to combine the results of low-fidelity (fast but inaccurate) and high-fidelity (slow but accurate) simulations. Despite the fast developments of multi-fidelity fusion techniques, most existing methods require particular data structures and do not scale well to high-dimensional output. To resolve these issues, we generalize the classic autoregression (AR), which is widely used due to its simplicity, robustness, accuracy, and tractability, and propose generalized autoregression (GAR) using tensor formulation and latent features. GAR can deal with arbitrary dimensional outputs and arbitrary multi-fidelity data structure to satisfy the demand of multi-fidelity fusion for complex problems; it admits a fully tractable likelihood and posterior requiring no approximate inference and scales well to high-dimensional problems. Furthermore, we prove the autokrigability theorem based on GAR in the multi-fidelity case and develop CIGAR, a simplified GAR with the same predictive mean accuracy but requires significantly less computation. In experiments of canonical PDEs and scientific computational examples, the proposed method consistently outperforms the SOTA methods with a large margin (up to 6x improvement in RMSE) with only a few high-fidelity training samples.

## [Linear tree shap](#)

- pengyu · Albert Bifet · Jesse Read · Chao Xu
- abstract@[open-review](#): Decision trees are well-known due to their ease of interpretability. To improve accuracy, we need to grow deep trees or ensembles of trees. These are hard to interpret, offsetting their original benefits. Shapley values have recently become a popular way to explain the predictions of tree-based machine learning models. It provides a linear weighting to features independent of the tree structure. The rise in popularity is mainly due to TreeShap, which solves a general exponential complexity problem in polynomial time. Following extensive adoption in the industry, more efficient algorithms are required. This paper presents a more efficient and straightforward algorithm: Linear TreeShap. Like TreeShap, Linear TreeShap is exact and requires the same amount of memory.

## [Decision Trees with Short Explainable Rules](#)

- Ferdinando Cicalese · Victor Feitosa Souza · Eduardo Laber · Marco Molinaro
- abstract@[open-review](#): Decision trees are widely used in many settings where interpretable models are preferred or required. As confirmed by recent empirical studies, the interpretability/explainability of a decision tree critically depends on some of its structural parameters, like size and the average/maximum depth of its leaves. There is indeed a vast literature on the design and analysis of decision tree algorithms that aim at optimizing these parameters. This paper contributes to this important line of research: we propose as a novel criterion of measuring the interpretability of a decision tree, the sparsity of the set of attributes that are (on average) required to explain the classification of the examples. We give a tight characterization of the best possible guarantees achievable by a decision tree built to optimize both our new measure (which we call the  $\text{explanation size}$ ) and the more classical measures of worst-case and average depth. In particular, we give an algorithm that guarantees  $\mathcal{O}(\ln n)$ -approximation (hence optimal if  $P \neq NP$ ) for the minimization of both the average/worst-case explanation size and the average/worst-case depth. In addition to our theoretical contributions, experiments with 20 real datasets show that our algorithm has accuracy competitive with CART while producing trees that allow for much simpler explanations.

## [Does Momentum Change the Implicit Regularization on Separable Data?](#)

- Bohan Wang · Qi Meng · Huishuai Zhang · Ruoyu Sun · Wei Chen · Zhi-Ming Ma · Tie-Yan Liu
- abstract@[open-review](#): The momentum acceleration technique is widely adopted in many optimization algorithms. However, there is no theoretical answer on how the momentum affects the generalization performance of the optimization algorithms. This paper studies this problem by analyzing the implicit regularization of momentum-based optimization. We prove that on the linear classification problem with separable data and exponential-tailed loss, gradient descent with momentum (GDM) converges to the  $L^2$  max-margin solution, which is the same as vanilla gradient descent. That means gradient descent with momentum acceleration still converges to a low-complexity model, which guarantees their generalization. We then analyze the stochastic and adaptive variants of GDM (i.e., SGDM and deterministic Adam) and show they also converge to the  $L^2$  max-margin solution.

Technically, to overcome the difficulty of the error accumulation in analyzing the momentum, we construct new potential functions to analyze the gap between the model parameter and the max-margin solution. Numerical experiments are conducted to support our theoretical results.

## [Efficient learning of nonlinear prediction models with time-series privileged information](#)

- Bastian Jung · Fredrik Johansson
- abstract@[open-review](#): In domains where sample sizes are limited, efficient learning algorithms are critical. Learning using privileged information (LuPI) offers increased sample efficiency by allowing prediction models access to types of information at training time which is unavailable when the models are used. In recent work, it was shown that for prediction in linear-Gaussian dynamical systems, a LuPI learner with access to intermediate time series data is never worse and often better in expectation than any unbiased classical learner. We provide new insights into this analysis and generalize it to nonlinear prediction tasks in latent dynamical systems, extending theoretical guarantees to the case where the map connecting latent variables and observations is known up to a linear transform. In addition, we propose algorithms based on random features and representation learning for the case when this map is unknown. A suite of empirical results confirm theoretical findings and show the potential of using privileged time-series information in nonlinear prediction.

## [Self-supervised Heterogeneous Graph Pre-training Based on Structural Clustering](#)

- Yaming Yang · Ziyu Guan · Zhe Wang · Wei Zhao · Cai Xu · Weigang Lu · Jianbin Huang
- abstract@[open-review](#): Recent self-supervised pre-training methods on Heterogeneous Information Networks (HINs) have shown promising competitiveness over traditional semi-supervised Heterogeneous Graph Neural Networks (HGNNs). Unfortunately, their performance heavily depends on careful customization of various strategies for generating high-quality positive examples and negative examples, which notably limits their flexibility and generalization ability. In this work, we present SHGP, a novel Self-supervised Heterogeneous Graph Pre-training approach, which does not need to generate any positive examples or negative examples. It consists of two modules that share the same attention-aggregation scheme. In each iteration, the Att-LPA module produces pseudo-labels through structural clustering, which serve as the self-supervision signals to guide the Att-HGNN module to learn object embeddings and attention coefficients. The two modules can effectively utilize and enhance each other, promoting the model to learn discriminative embeddings. Extensive experiments on four real-world datasets demonstrate the superior effectiveness of SHGP against state-of-the-art unsupervised baselines and even semi-supervised baselines. We will release our source code at GitHub once the manuscript is accepted.

## [Object Scene Representation Transformer](#)

- Mehdi S. M. Sajjadi · Daniel Duckworth · Aravindh Mahendran · Sjoerd van Steenkiste · Filip Pavetić · Mario Lucic · Leonidas Guibas · Klaus Greff · Thomas Kipf
- abstract@[open-review](#): A compositional understanding of the world in terms of objects and their geometry in 3D space is considered a cornerstone of human cognition. Facilitating the learning of such a representation in neural networks holds promise for substantially improving labeled data efficiency. As a key step in this direction, we make progress on the problem of learning 3D-consistent decompositions of complex scenes into individual objects in an unsupervised fashion. We introduce Object Scene Representation Transformer (OSRT), a 3D-centric model in which individual object representations naturally emerge through novel view synthesis. OSRT scales to significantly more complex scenes with larger diversity of objects and backgrounds than existing methods. At the same time, it is multiple orders of magnitude faster at compositional rendering thanks to its light field parametrization and the novel Slot Mixer decoder. We believe this work will not only accelerate future architecture exploration and scaling efforts, but it will also serve as a useful tool for both object-centric as well as neural scene representation learning communities.

## [PointNeXt: Revisiting PointNet++ with Improved Training and Scaling Strategies](#)

- Guocheng Qian · Yuchen Li · Houwen Peng · Jinjie Mai · Hasan Hammoud · Mohamed Elhoseiny · Bernard Ghanem
- abstract@[open-review](#): PointNet++ is one of the most influential neural architectures for point cloud understanding. Although the accuracy of PointNet++ has been largely surpassed by recent networks such as PointMLP and Point Transformer, we find that a large portion of the performance gain is due to improved training strategies, \textit{i.e.} data augmentation and optimization techniques, and increased model sizes rather than architectural innovations. Thus, the full potential of PointNet++ has yet to be explored. In this work, we revisit the classical PointNet++ through a systematic study of model training and scaling strategies, and offer two major contributions. First, we propose a set of improved training strategies that significantly boost the performance of PointNet++. For example, we show that, without any change in architecture, PointNet++ can even achieve slightly higher accuracy than the state-of-the-art PointMLP on the ScanObjectNN benchmark. Second, we introduce an inverted residual bottleneck design and separable MLPs into PointNet++ to enable effective and efficient model scaling and propose \textit{PointNeXt}, a scalable version of PointNets. PointNeXt can be flexibly scaled up and outperforms state-of-the-art methods on both 3D classification and segmentation tasks. For classification, PointNeXt reaches an overall accuracy of 87.7% on ScanObjectNN, surpassing PointMLP by 2.3%, while being 10 times faster in inference. For semantic segmentation, PointNeXt establishes a new state-of-the-art performance with 74.9% mean IoU on S3DIS (6-fold cross-validation), being superior to the recent Point Transformer. The code and models will be made publicly available.

## [Explicable Policy Search](#)

- Ze Gong · Yu ("Tony") Zhang
- abstract@[open-review](#): Human teammates often form conscious and subconscious expectations of each other to better interact. Teaming success is contingent on whether such expectations can be met. For an intelligent agent to operate with humans beside them, similarly, it must consider the human's expectation of its behavior. Otherwise, it may result in loss of trust and degraded team performance. A key challenge here is that the human's expectation may not align with the agent's optimal behavior, due to, for example, the human's partial or inaccurate understanding of the task domain. Prior work on explicable planning describes the ability of agents to respect their human teammate's expectations by trading off task performance for more expected or "explicable" behaviors. In this paper, we introduce Explicable Policy Search (EPS) to significantly extend such an ability to a reinforcement learning (RL) setting and to handle stochastic domains with continuous state and action spaces. However, in contrast to the traditional RL method, EPS must at the same time infer the human's hidden expectations. Such inferences require information about the human's belief about domain dynamics and her reward model but directly querying them is impractical. We demonstrate that they can be sufficiently encoded by a surrogate reward function, which can be learned based on the human's feedback on the agent's behavior. The surrogate reward function is then used to reshape the agent's reward function, which is shown to be equivalent to searching for an explicable policy. We evaluate our method for EPS in a set of continuous navigation domains with synthetic human models and in an autonomous driving domain with a user study. The results suggest that our method can generate explicable behaviors that reconcile task performance with human expectation intelligently and has real-world relevance in human-agent teaming domains.

## [Module-Aware Optimization for Auxiliary Learning](#)

- Hong Chen · Xin Wang · Yue Liu · Yuwei Zhou · Chaoyu Guan · Wenwu Zhu
- abstract@[open-review](#): Auxiliary learning is a widely adopted practice in deep learning, which aims to improve the model performance on the primary task by exploiting the beneficial information in the auxiliary loss. Existing auxiliary learning methods only focus on balancing the auxiliary loss and the primary loss, ignoring the module-level auxiliary influence, i.e., an auxiliary loss will be beneficial for optimizing specific modules within the model but harmful to others, failing to make full use of auxiliary information. To tackle the problem, we propose a Module-Aware Optimization approach for Auxiliary Learning (MAOAL). The proposed approach considers the module-level influence through the learnable module-level auxiliary importance, i.e.,

the importance of each auxiliary loss to each module. Specifically, the proposed approach jointly optimizes the module-level auxiliary importance and the model parameters in a bi-level manner. In the lower optimization, the model parameters are optimized with the importance parameterized gradient, while in the upper optimization, the module-level auxiliary importance is updated with the implicit gradient from a small developing dataset. Extensive experiments show that our proposed MAOAL method consistently outperforms state-of-the-art baselines for different auxiliary losses on various datasets, demonstrating that our method can serve as a powerful generic tool for auxiliary learning.

## [Model-Based Imitation Learning for Urban Driving](#)

- Anthony Hu · Gianluca Corrado · Nicolas Griffiths · Zachary Murez · Corina Gurau · Hudson Yeo · Alex Kendall · Roberto Cipolla · Jamie Shotton
- abstract@[open-review](#): An accurate model of the environment and the dynamic agents acting in it offers great potential for improving motion planning. So far, such world models have been shown to be highly effective at solving games, but only in simple visual environments with little interaction among agents. We present MILE: a Model-based Imitation LEarning approach for autonomous driving that scales to the complexity of urban driving scenes. Our approach leverages 3D geometry as an inductive bias and learns a highly compact latent space directly from high resolution videos of expert demonstrations. MILE learns a model of the world and a driving policy from an offline corpus of driving data, without any online interaction with the environment. Our method improves upon prior state-of-the-art by 35% in driving score on the CARLA simulator when deployed in a completely new town and new weather conditions. Further, we qualitatively show that our model can predict diverse and plausible future scenes in bird's-eye view over a long time horizon (>60s), that are consistent with predicted ego-actions.

## [Self-Supervised Learning with an Information Maximization Criterion](#)

- Serdar Ozsoy · Shadi Hamdan · Sercan Arik · Deniz Yuret · Alper Erdogan
- abstract@[open-review](#): Self-supervised learning allows AI systems to learn effective representations from large amounts of data using tasks that do not require costly labeling. Mode collapse, i.e., the model producing identical representations for all inputs, is a central problem to many self-supervised learning approaches, making self-supervised tasks, such as matching distorted variants of the inputs, ineffective. In this article, we argue that a straightforward application of information maximization among alternative latent representations of the same input naturally solves the collapse problem and achieves competitive empirical results. We propose a self-supervised learning method, CorInfoMax, that uses a second-order statistics-based mutual information measure that reflects the level of correlation among its arguments. Maximizing this correlative information measure between alternative representations of the same input serves two purposes: (1) it avoids the collapse problem by generating feature vectors with non-degenerate covariances; (2) it establishes relevance among alternative representations by increasing the linear dependence among them. An approximation of the proposed information maximization objective simplifies to a Euclidean distance-based objective function regularized by the log-determinant of the feature covariance matrix. The regularization term acts as a natural barrier against feature space degeneracy. Consequently, beyond avoiding complete output collapse to a single point, the proposed approach also prevents dimensional collapse by encouraging the spread of information across the whole feature space. Numerical experiments demonstrate that CorInfoMax achieves better or competitive performance results relative to the state-of-the-art SSL approaches.

## [LogiGAN: Learning Logical Reasoning via Adversarial Pre-training](#)

- Xinyu Pi · Wanjun Zhong · Yan Gao · Nan Duan · Jian-Guang Lou
- abstract@[open-review](#): We present LogiGAN, an unsupervised adversarial pre-training framework for improving logical reasoning abilities of language models. Upon automatic identification of logical reasoning phenomena in massive text corpus via detection heuristics, we train language models to predict the masked-out logical statements. Inspired by the facilitation effect of reflective thinking in human learning, we analogically simulate the learning-thinking process with an adversarial Generator-Verifier architecture to assist logic learning. LogiGAN implements a novel sequential GAN approach that (a) circumvents the non-differentiable challenge of the sequential GAN by leveraging the Generator as a sentence-level generative likelihood scorer with a learning objective of reaching scoring consensus with the Verifier; (b) is computationally feasible for large-scale pre-training with arbitrary target length. Both base and large size language models pre-trained with LogiGAN demonstrate obvious performance improvement on 12 datasets requiring general reasoning abilities, revealing the fundamental role of logic in broad reasoning, as well as the effectiveness of LogiGAN. Ablation studies on LogiGAN components reveal the relative orthogonality between linguistic and logic abilities and suggest that reflective thinking's facilitation effect might also generalize to machine learning.

## [Revisiting Realistic Test-Time Training: Sequential Inference and Adaptation by Anchored Clustering](#)

- Yongyi Su · Xun Xu · Kui Jia
- abstract@[open-review](#): Deploying models on target domain data subject to distribution shift requires adaptation. Test-time training (TTT) emerges as a solution to this adaptation under a realistic scenario where access to full source domain data is not available and instant inference on target domain is required. Despite many efforts into TTT, there is a confusion over the experimental settings, thus leading to unfair comparisons. In this work, we first revisit TTT assumptions and categorize TTT protocols by two key factors. Among the multiple protocols, we adopt a realistic sequential test-time training (sTTT) protocol, under which we further develop a test-time anchored clustering (TTAC) approach to enable stronger test-time feature learning. TTAC discovers clusters in both source and target domain and match the target clusters to the source ones to improve generalization. Pseudo label filtering and iterative updating are developed to improve the effectiveness and efficiency of anchored clustering. We demonstrate that under all TTT protocols TTAC consistently outperforms the state-of-the-art methods on five TTT datasets. We hope this work will provide a fair benchmarking of TTT methods and future research should be compared within respective protocols.

## [TokenMixup: Efficient Attention-guided Token-level Data Augmentation for Transformers](#)

- Hyeong Kyu Choi · Joonmyung Choi · Hyunwoo Kim
- abstract@[open-review](#): Mixup is a commonly adopted data augmentation technique for image classification. Recent advances in mixup methods primarily focus on mixing based on saliency. However, many saliency detectors require intense computation and are especially burdensome for parameter-heavy transformer models. To this end, we propose TokenMixup, an efficient attention-guided token-level data augmentation method that aims to maximize the saliency of a mixed set of tokens. TokenMixup provides ~15 faster saliency-aware data augmentation compared to gradient-based mixup methods. Moreover, we introduce a variant of TokenMixup which mixes tokens within a single instance, thereby enabling multi-scale feature augmentation. Experiments show that our methods significantly improve the baseline models' performance on CIFAR and ImageNet-1K, while being more efficient than previous methods. We also reach state-of-the-art performance on CIFAR-100 among from-scratch transformer models. Code will be released.

## [Synthetic Model Combination: An Instance-wise Approach to Unsupervised Ensemble Learning](#)

- Alex Chan · Mihaela van der Schaar
- abstract@[open-review](#): Consider making a prediction over new test data without any opportunity to learn from a training set of labelled data - instead given access to a set of expert models and their predictions alongside some limited information about the dataset used to train them. In scenarios from finance to the medical sciences, and even consumer practice, stakeholders have developed models on private data they either cannot, or do not want to, share. Given the value and legislation surrounding personal information, it is not surprising that only the models, and not the data, will be released - the pertinent question becoming: how best to use these models? Previous work has focused on global model selection or ensembling, with the result of a

single final model across the feature space. Machine learning models perform notoriously poorly on data outside their training domain however, and so we argue that when ensembling models the weightings for individual instances must reflect their respective domains - in other words models that are more likely to have seen information on that instance should have more attention paid to them. We introduce a method for such an instance-wise ensembling of models, including a novel representation learning step for handling sparse high-dimensional domains. Finally, we demonstrate the need and generalisability of our method on classical machine learning tasks as well as highlighting a real world use case in the pharmacological setting of vancomycin precision dosing.

## [4D Unsupervised Object Discovery](#)

- Yuqi Wang · Yuntao Chen · ZHAO-XIANG ZHANG
- abstract@[open-review](#): Object discovery is a core task in computer vision. While tremendous success has progressed in supervised object detection with vast annotated data, its unsupervised counterpart remains largely unexplored. With the growth of data volume, the expensive cost of annotations is the major limitation hindering further study. Therefore, discovering the objects without annotations has great significance. However, the task seems impractical on still-image or point cloud alone due to the lack of discriminative information. Previous studies overlook the crucial temporal information and constraints naturally behind multi-modality. In this paper, we propose 4D unsupervised object discovery, jointly discovering objects from 4D data—3D point clouds and 2D RGB images with temporal information. We present the first practical approach for this task by proposing a ClusterNet on 3D point clouds, which is joint iterative optimizing with a 2D localization network. Extensive experiments on the large-scale Waymo Open Dataset suggest that the localization network and ClusterNet achieve competitive performance on class-agnostic 2D object detection and 3D instance segmentation, bridging the gap between unsupervised methods and full supervision. Code will be released.

## [Optimistic Tree Searches for Combinatorial Black-Box Optimization](#)

- Cedric Malherbe · Antoine Grosnit · Rasul Tutunov · Haitham Bou Ammar · Jun Wang
- abstract@[open-review](#): The optimization of combinatorial black-box functions is pervasive in computer science and engineering. However, the combinatorial explosion of the search space and lack of natural ordering pose significant challenges for current techniques from a theoretical and practical perspective, and require new algorithmic ideas. In this paper, we propose to adapt the recent advances in tree searches and partitioning techniques to design and analyze novel black-box combinatorial solvers. A first contribution is the analysis of a first tree-search algorithm called Optimistic Lipschitz Tree Search (OLTS) which assumes the Lipschitz constant of the function to be known. Linear convergence rates are provided for this algorithm under specific conditions, improving upon the logarithmic rates of baselines. An adaptive version, called Optimistic Combinatorial Tree Search (OCTS), is then introduced for the more realistic setup where we do not have any information on the Lipschitz constant of the function. Similar theoretical guarantees are shown to hold for OCTS and a numerical assessment is provided to illustrate the potential of tree searches with respect to state-of-the-art methods over typical benchmarks.

## [Learning robust rule representations for abstract reasoning via internal inferences](#)

- Wenbo Zhang · likai tang · Site Mo · Xianggen Liu · Sen Song
- abstract@[open-review](#): Abstract reasoning, as one of the hallmarks of human intelligence, involves collecting information, identifying abstract rules, and applying the rules to solve new problems. Although the neural networks have achieved human-level performances in several tasks, the abstract reasoning techniques still far lag behind due to the complexity of learning and applying the logic rules, especially in an unsupervised manner. In this work, we propose a novel framework, ARII, that learns rule representations for Abstract Reasoning via Internal Inferences. The key idea is to repeatedly apply a rule to different instances in hope of having a comprehensive understanding (i.e., representations) of the rule. Specifically, ARII consists of a rule encoder, a reasoner, and an internal referrer. Based on the representations produced by the rule encoder, the reasoner draws the conclusion while the referrer performs internal inferences to regularize rule representations to be robust and generalizable. We evaluate ARII on two benchmark datasets, including PGM and I-RAVEN. We observe that ARII achieves the new state-of-the-art records on the majority of the reasoning tasks, including most of the generalization tests in PGM.

## [GBA: A Tuning-free Approach to Switch between Synchronous and Asynchronous Training for Recommendation Models](#)

- Wenbo Su · Yuanxing Zhang · Yufeng Cai · Kaixu Ren · Pengjie Wang · Huimin Yi · Yue Song · Jing Chen · Hongbo Deng · Jian Xu · Lin Qu · Bo Zheng
- abstract@[open-review](#): High-concurrency asynchronous training upon parameter server (PS) architecture and high-performance synchronous training upon all-reduce (AR) architecture are the most commonly deployed distributed training modes for recommender systems. Although the synchronous AR training is designed to have higher training efficiency, the asynchronous PS training would be a better choice on training speed when there are stragglers (slow workers) in the shared cluster, especially under limited computing resources. To take full advantages of these two training modes, an ideal way is to switch between them upon the cluster status. We find two obstacles to a tuning-free approach: the different distribution of the gradient values and the stale gradients from the stragglers. In this paper, we propose Global Batch gradients Aggregation (GBA) over PS, which aggregates and applies gradients with the same global batch size as the synchronous training. A token-control process is implemented to assemble the gradients and decay the gradients with severe staleness. We provide the convergence analysis to reveal that GBA has comparable convergence properties with the synchronous training, and demonstrate the robustness of GBA the recommendation models against the gradient staleness. Experiments on three industrial-scale recommendation tasks show that GBA is an effective tuning-free approach for switching. Compared to the state-of-the-art derived asynchronous training, GBA achieves up to 0.2% improvement on the AUC metric, which is significant for the recommendation models. Meanwhile, under the strained hardware resource, GBA speeds up at least 2.4x compared to the synchronous training.

## [Adversarial Style Augmentation for Domain Generalized Urban-Scene Segmentation](#)

- Zhun Zhong · Yuyang Zhao · Gim Hee Lee · Nicu Sebe
- abstract@[open-review](#): In this paper, we consider the problem of domain generalization in semantic segmentation, which aims to learn a robust model using only labeled synthetic (source) data. The model is expected to perform well on unseen real (target) domains. Our study finds that the image style variation can largely influence the model's performance and the style features can be well represented by the channel-wise mean and standard deviation of images. Inspired by this, we propose a novel adversarial style augmentation (AdvStyle) approach, which can dynamically generate hard stylized images during training and thus can effectively prevent the model from overfitting on the source domain. Specifically, AdvStyle regards the style feature as a learnable parameter and updates it by adversarial training. The learned adversarial style feature is used to construct an adversarial image for robust model training. AdvStyle is easy to implement and can be readily applied to different models. Experiments on two synthetic-to-real semantic segmentation benchmarks demonstrate that AdvStyle can significantly improve the model performance on unseen real domains and show that we can achieve the state of the art. Moreover, AdvStyle can be employed to domain generalized image classification and produces a clear improvement on the considered datasets.

## [Learning Graph-embedded Key-event Back-tracing for Object Tracking in Event Clouds](#)

- Zhiyu Zhu · Junhui Hou · Xianqiang Lyu
- abstract@[open-review](#): Event data-based object tracking is attracting attention increasingly. Unfortunately, the unusual data structure caused by the unique sensing mechanism poses great challenges in designing downstream algorithms. To tackle such challenges, existing methods usually re-organize raw event data (or event clouds) with the event frame/image representation to adapt to mature RGB data-based tracking paradigms, which compromises the high

temporal resolution and sparse characteristics. By contrast, we advocate developing new designs/techniques tailored to the special data structure to realize object tracking. To this end, we make the first attempt to construct a new end-to-end learning-based paradigm that directly consumes event clouds. Specifically, to process a non-uniformly distributed large-scale event cloud efficiently, we propose a simple yet effective density-insensitive downsampling strategy to sample a subset called key-events. Then, we employ a graph-based network to embed the irregular spatio-temporal information of key-events into a high-dimensional feature space, and the resulting embeddings are utilized to predict their target likelihoods via semantic-driven Siamese-matching. Besides, we also propose motion-aware target likelihood prediction, which learns the motion flow to back-trace the potential initial positions of key-events and measures them with the previous proposal. Finally, we obtain the bounding box by adaptively fusing the two intermediate ones separately regressed from the weighted embeddings of key-events by the two types of predicted target likelihoods. Extensive experiments on both synthetic and real event datasets demonstrate the superiority of the proposed framework over state-of-the-art methods in terms of both the tracking accuracy and speed. The code is publicly available at <https://github.com/ZHU-Zhiyu/Event-tracking>.

## [Self-Supervised Aggregation of Diverse Experts for Test-Agnostic Long-Tailed Recognition](#)

- Yifan Zhang · Bryan Hooi · Lanqing Hong · Jiashi Feng
- abstract@[open-review](#): Existing long-tailed recognition methods, aiming to train class-balanced models from long-tailed data, generally assume the models would be evaluated on the uniform test class distribution. However, practical test class distributions often violate this assumption (e.g., being either long-tailed or even inversely long-tailed), which may lead existing methods to fail in real applications. In this paper, we study a more practical yet challenging task, called test-agnostic long-tailed recognition, where the training class distribution is long-tailed while the test class distribution is agnostic and not necessarily uniform. In addition to the issue of class imbalance, this task poses another challenge: the class distribution shift between the training and test data is unknown. To tackle this task, we propose a novel approach, called Self-supervised Aggregation of Diverse Experts, which consists of two strategies: (i) a new skill-diverse expert learning strategy that trains multiple experts from a single and stationary long-tailed dataset to separately handle different class distributions; (ii) a novel test-time expert aggregation strategy that leverages self-supervision to aggregate the learned multiple experts for handling unknown test class distributions. We theoretically show that our self-supervised strategy has a provable ability to simulate test-agnostic class distributions. Promising empirical results demonstrate the effectiveness of our method on both vanilla and test-agnostic long-tailed recognition. Source code is available in the supplementary material.

## [Amortized Projection Optimization for Sliced Wasserstein Generative Models](#)

- Khai Nguyen · Nhat Ho
- abstract@[open-review](#): Seeking informative projecting directions has been an important task in utilizing sliced Wasserstein distance in applications. However, finding these directions usually requires an iterative optimization procedure over the space of projecting directions, which is computationally expensive. Moreover, the computational issue is even more severe in deep learning applications, where computing the distance between two mini-batch probability measures is repeated several times. This nested-loop has been one of the main challenges that prevent the usage of sliced Wasserstein distances based on good projections in practice. To address this challenge, we propose to utilize the \textit{learning-to-optimize} technique or \textit{amortized optimization} to predict the informative direction of any given two mini-batch probability measures. To the best of our knowledge, this is the first work that bridges amortized optimization and sliced Wasserstein generative models. In particular, we derive linear amortized models, generalized linear amortized models, and non-linear amortized models which are corresponding to three types of novel mini-batch losses, named \emph{amortized sliced Wasserstein}. We demonstrate the favorable performance of the proposed sliced losses in deep generative modeling on standard benchmark datasets.

## [OpenAUC: Towards AUC-Oriented Open-Set Recognition](#)

- Zitai Wang · Qianqian Xu · Zhiyong Yang · Yuan He · Xiaochun Cao · Qingming Huang
- abstract@[open-review](#): Traditional machine learning follows a close-set assumption that the training and test set share the same label space. While in many practical scenarios, it is inevitable that some test samples belong to unknown classes (open-set). To fix this issue, Open-Set Recognition (OSR), whose goal is to make correct predictions on both close-set samples and open-set samples, has attracted rising attention. In this direction, the vast majority of literature focuses on the pattern of open-set samples. However, how to evaluate model performance in this challenging task is still unsolved. In this paper, a systematic analysis reveals that most existing metrics are essentially inconsistent with the aforementioned goal of OSR: (1) For metrics extended from close-set classification, such as Open-Set F-score, Youden's index, and Normalized Accuracy, a poor open-set prediction can escape from a low performance score with a superior close-set prediction. (2) Novelty detection AUC, which measures the ranking performance between close-set and open-set samples, ignores the close-set performance. To fix these issues, we propose a novel metric named OpenAUC. Compared with existing metrics, OpenAUC enjoys a concise pairwise formulation that evaluates open-set performance and close-set performance in a coupling manner. On top of this, further analysis shows that OpenAUC is free from the inconsistency properties of existing metrics. Finally, an end-to-end learning method is proposed to minimize the OpenAUC risk, and the experimental results on popular benchmark datasets speak to its effectiveness.

## [DeepTOP: Deep Threshold-Optimal Policy for MDPs and RMBAs](#)

- Khaled Nakhleh · I-Hong Hou
- abstract@[open-review](#): We consider the problem of learning the optimal threshold policy for control problems. Threshold policies make control decisions by evaluating whether an element of the system state exceeds a certain threshold, whose value is determined by other elements of the system state. By leveraging the monotone property of threshold policies, we prove that their policy gradients have a surprisingly simple expression. We use this simple expression to build an off-policy actor-critic algorithm for learning the optimal threshold policy. Simulation results show that our policy significantly outperforms other reinforcement learning algorithms due to its ability to exploit the monotone property. In addition, we show that the Whittle index, a powerful tool for restless multi-armed bandit problems, is equivalent to the optimal threshold policy for an alternative problem. This observation leads to a simple algorithm that finds the Whittle index by learning the optimal threshold policy in the alternative problem. Simulation results show that our algorithm learns the Whittle index much faster than several recent studies that learn the Whittle index through indirect means.

## [Don't Pour Cereal into Coffee: Differentiable Temporal Logic for Temporal Action Segmentation](#)

- Ziwei Xu · Yogesh Rawat · Yongkang Wong · Mohan Kankanhalli · Mubarak Shah
- abstract@[open-review](#): We propose Differentiable Temporal Logic (DTL), a model-agnostic framework that introduces temporal constraints to deep networks. DTL treats the outputs of a network as a truth assignment of a temporal logic formula, and computes a temporal logic loss reflecting the consistency between the output and the constraints. We propose a comprehensive set of constraints, which are implicit in data annotations, and incorporate them with deep networks via DTL. We evaluate the effectiveness of DTL on the temporal action segmentation task and observe improved performance and reduced logical errors in the output of different task models. Furthermore, we provide an extensive analysis to visualize the desirable effects of DTL.

## [Distributionally Robust Optimization with Data Geometry](#)

- Jiashuo Liu · Jiayun Wu · Bo Li · Peng Cui
- abstract@[open-review](#): Distributionally Robust Optimization (DRO) serves as a robust alternative to empirical risk minimization (ERM), which optimizes the worst-case distribution in an uncertainty set typically specified by distance metrics including  $\ell_p$ -divergence and the Wasserstein distance. The metrics defined in the ostensible high dimensional space lead to exceedingly large uncertainty sets, resulting in the underperformance of most existing DRO methods. It has been well documented that high dimensional data approximately resides on low dimensional manifolds. In this work, to further constrain

the uncertainty set, we incorporate data geometric properties into the design of distance metrics, obtaining our novel Geometric Wasserstein DRO (GDRO). Empowered by Gradient Flow, we derive a generically applicable approximate algorithm for the optimization of GDRO, and provide the bounded error rate of the approximation as well as the convergence rate of our algorithm. We also theoretically characterize the edge cases where certain existing DRO methods are the degeneracy of GDRO. Extensive experiments justify the superiority of our GDRO to existing DRO methods in multiple settings with strong distributional shifts, and confirm that the uncertainty set of GDRO adapts to data geometry.

## [Revisiting Sliced Wasserstein on Images: From Vectorization to Convolution](#)

- Khai Nguyen Â· Nhat Ho
- abstract@[open-review](#): The conventional sliced Wasserstein is defined between two probability measures that have realizations as  $\text{vectors}$ . When comparing two probability measures over images, practitioners first need to vectorize images and then project them to one-dimensional space by using matrix multiplication between the sample matrix and the projection matrix. After that, the sliced Wasserstein is evaluated by averaging the two corresponding one-dimensional projected probability measures. However, this approach has two limitations. The first limitation is that the spatial structure of images is not captured efficiently by the vectorization step; therefore, the later slicing process becomes harder to gather the discrepancy information. The second limitation is memory inefficiency since each slicing direction is a vector that has the same dimension as the images. To address these limitations, we propose novel slicing methods for sliced Wasserstein between probability measures over images that are based on the convolution operators. We derive  $\text{convolution sliced Wasserstein}$  (CSW) and its variants via incorporating stride, dilation, and non-linear activation function into the convolution operators. We investigate the metricity of CSW as well as its sample complexity, its computational complexity, and its connection to conventional sliced Wasserstein distances. Finally, we demonstrate the favorable performance of CSW over the conventional sliced Wasserstein in comparing probability measures over images and in training deep generative modeling on images.

## [Rank Diminishing in Deep Neural Networks](#)

- Ruili Feng Â· Kecheng Zheng Â· Yukun Huang Â· Deli Zhao Â· Michael Jordan Â· Zheng-Jun Zha
- abstract@[open-review](#): The rank of neural networks measures information flowing across layers. It is an instance of a key structural condition that applies across broad domains of machine learning. In particular, the assumption of low-rank feature representations led to algorithmic developments in many architectures. For neural networks, however, the intrinsic mechanism that yields low-rank structures remains vague and unclear. To fill this gap, we perform a rigorous study on the behavior of network rank, focusing particularly on the notion of rank deficiency. We theoretically establish a universal monotone decreasing property of network ranks from the basic rules of differential and algebraic composition, and uncover rank deficiency of network blocks and deep function coupling. By virtue of our numerical tools, we provide the first empirical analysis of the per-layer behavior of network ranks in realistic settings,  $\text{VGG}$ , ResNets, deep MLPs, and Transformers on ImageNet. These empirical results are in direct accord with our theory. Furthermore, we reveal a novel phenomenon of independence deficit caused by the rank deficiency of deep networks, where classification confidence of a given category can be linearly decided by the confidence of a handful of other categories. The theoretical results of this work, together with the empirical findings, may advance understanding of the inherent principles of deep neural networks.

## [A Lower Bound of Hash Codes' Performance](#)

- Xiaosu Zhu Â· Jingkuan Song Â· Yu Lei Â· Lianli Gao Â· Hengtao Shen
- abstract@[open-review](#): As a crucial approach for compact representation learning, hashing has achieved great success in effectiveness and efficiency. Numerous heuristic Hamming space metric learning objectives are designed to obtain high-quality hash codes. Nevertheless, a theoretical analysis of criteria for learning good hash codes remains largely unexploited. In this paper, we prove that inter-class distinctiveness and intra-class compactness among hash codes determine the lower bound of hash codes' performance. Promoting these two characteristics could lift the bound and improve hash learning. We then propose a surrogate model to fully exploit the above objective by estimating the posterior of hash codes and controlling it, which results in a low-bias optimization. Extensive experiments reveal the effectiveness of the proposed method. By testing on a series of hash-models, we obtain performance improvements among all of them, with an up to \$26.5\%\$ increase in mean Average Precision and an up to \$20.5\%\$ increase in accuracy.

## [I2Q: A Fully Decentralized Q-Learning Algorithm](#)

- Jiechuan Jiang Â· Zongqing Lu
- abstract@[open-review](#): Fully decentralized multi-agent reinforcement learning has shown great potentials for many real-world cooperative tasks, where the global information,  $\text{e.g.}$ , the actions of other agents, is not accessible. Although independent Q-learning is widely used for decentralized training, the transition probabilities are non-stationary since other agents are updating policies simultaneously, which leads to non-guaranteed convergence of independent Q-learning. To deal with non-stationarity, we first introduce stationary ideal transition probabilities, on which independent Q-learning could converge to the global optimum. Further, we propose a fully decentralized method, I2Q, which performs independent Q-learning on the modeled ideal transition function to reach the global optimum. The modeling of ideal transition function in I2Q is fully decentralized and independent from the learned policies of other agents, helping I2Q be free from non-stationarity and learn the optimal policy. Empirically, we show that I2Q can achieve remarkable improvement in a variety of cooperative multi-agent tasks.

## [Unifying Voxel-based Representation with Transformer for 3D Object Detection](#)

- Yanwei Li Â· Yilun Chen Â· Xiaojuan Qi Â· Zeming Li Â· Jian Sun Â· Jiaya Jia
- abstract@[open-review](#): In this work, we present a unified framework for multi-modality 3D object detection, named UVTR. The proposed method aims to unify multi-modality representations in the voxel space for accurate and robust single- or cross-modality 3D detection. To this end, the modality-specific space is first designed to represent different inputs in the voxel feature space. Different from previous work, our approach preserves the voxel space without height compression to alleviate semantic ambiguity and enable spatial interactions. Benefit from the unified manner, cross-modality interaction is then proposed to make full use of inherent properties from different sensors, including knowledge transfer and modality fusion. In this way, geometry-aware expressions in point clouds and context-rich features in images are well utilized for better performance and robustness. The transformer decoder is applied to efficiently sample features from the unified space with learnable positions, which facilitates object-level interactions. In general, UVTR presents an early attempt to represent different modalities in a unified framework. It surpasses previous work in single- and multi-modality entries and achieves leading performance in the nuScenes test set with 69.7%, 55.1%, and 71.1% NDS for LiDAR, camera, and multi-modality inputs, respectively. Our code will be made publicly available.

## [Multiple-sample Neural Image Compression](#)

- Tongda Xu Â· Yan Wang Â· Dailan He Â· Chenjian Gao Â· Han Gao Â· Kunzan Liu Â· Hongwei Qin
- abstract@[open-review](#): This paper considers the problem of lossy neural image compression (NIC). Current state-of-the-art (SOTA) methods adopt uniform posterior to approximate quantization noise, and single-sample pathwise estimator to approximate the gradient of evidence lower bound (ELBO). In this paper, we propose to train NIC with multiple-sample importance weighted autoencoder (IWAE) target, which is tighter than ELBO and converges to log likelihood as sample size increases. First, we identify that the uniform posterior of NIC has special properties, which affect the variance and bias of pathwise and score function estimators of the IWAE target. Moreover, we provide insights on a commonly adopted trick in NIC from gradient variance perspective. Based on those analysis, we further propose multiple-sample NIC (MS-NIC), an enhanced IWAE target for NIC. Experimental results demonstrate that it improves SOTA NIC methods. Our MS-NIC is plug-and-play, and can be easily extended to neural video compression.

## [The Unreliability of Explanations in Few-Shot In-Context Learning](#)

- Xi Ye · Greg Durrett
- abstract@[open-review](#): Does prompting a large language model like GPT-3 with explanations improve in-context learning? We study this question specifically on two NLP tasks that involve reasoning over text, namely question answering and natural language inference. For these tasks, we find that including explanations GPT-3's prompt and having the model generate them only mildly improves accuracy over standard few-shot learning, contrary to recent results on symbolic reasoning tasks. Moreover, explanations generated by GPT-3 may not entail the predictions nor be factually grounded in the input, even on simple tasks with extractive explanations. However, these flawed explanations can still be useful as a way to verify GPT-3's predictions post-hoc. Through analysis in three settings, we show that explanations judged as good by humans---those that are logically consistent with the input and the prediction---usually cooccur with more accurate predictions. Following these observations, we present a framework for calibrating model predictions based on the reliability of the explanations. We train calibrators using automatically extracted scores that approximately assess the reliability of explanations, which helps improve performance across three different datasets.

## [Regularized Gradient Descent Ascent for Two-Player Zero-Sum Markov Games](#)

- Sihan Zeng · Thinh Doan · Justin Romberg
- abstract@[open-review](#): We study the problem of finding the Nash equilibrium in a two-player zero-sum Markov game. Due to its formulation as a minimax optimization program, a natural approach to solve the problem is to perform gradient descent/ascent with respect to each player in an alternating fashion. However, due to the non-convexity/non-concavity of the underlying objective function, theoretical understandings of this method are limited. In our paper, we consider solving an entropy-regularized variant of the Markov game. The regularization introduces structures into the optimization landscape that make the solutions more identifiable and allow the problem to be solved more efficiently. Our main contribution is to show that under proper choices of the regularization parameter, the gradient descent ascent algorithm converges to the Nash equilibrium of the original unregularized problem. We explicitly characterize the finite-time performance of the last iterate of our algorithm, which vastly improves over the existing convergence bound of the gradient descent ascent algorithm without regularization. Finally, we complement the analysis with numerical simulations that illustrate the accelerated convergence of the algorithm.

## [The price of unfairness in linear bandits with biased feedback](#)

- Solenne Gaucher · Alexandra Carpentier · Christophe Giraud
- abstract@[open-review](#): In this paper, we study the problem of fair sequential decision making with biased linear bandit feedback. At each round, a player selects an action described by a covariate and by a sensitive attribute. The perceived reward is a linear combination of the covariates of the chosen action, but the player only observes a biased evaluation of this reward, depending on the sensitive attribute. To characterize the difficulty of this problem, we design a phased elimination algorithm that corrects the unfair evaluations, and establish upper bounds on its regret. We show that the worst-case regret is smaller than  $\mathcal{O}(\kappa^{1/3} \log(T)^{1/3} T^{2/3})$ , where  $\kappa$  is an explicit geometrical constant characterizing the difficulty of bias estimation. We prove lower bounds on the worst-case regret for some sets of actions showing that this rate is tight up to a possible sub-logarithmic factor. We also derive gap-dependent upper bounds on the regret, and matching lower bounds for some problem instance. Interestingly, these results reveal a transition between a regime where the problem is as difficult as its unbiased counterpart, and a regime where it can be much harder.

## [TransTab: Learning Transferable Tabular Transformers Across Tables](#)

- Zifeng Wang · Jimeng Sun
- abstract@[open-review](#): Tabular data (or tables) are the most widely used data format in machine learning (ML). However, ML models often assume the table structure keeps fixed in training and testing. Before ML modeling, heavy data cleaning is required to merge disparate tables with different columns. This preprocessing often incurs significant data waste (e.g., removing unmatched columns and samples). How to learn ML models from multiple tables with partially overlapping columns? How to incrementally update ML models as more columns become available over time? Can we leverage model pretraining on multiple distinct tables? How to train an ML model which can predict on an unseen table? To answer all those questions, we propose to relax fixed table structures by introducing a Transferable Tabular Transformer (TransTab) for tables. The goal of TransTab is to convert each sample (a row in the table) to a generalizable embedding vector, and then apply stacked transformers for feature encoding. One methodology insight is combining column description and table cells as the raw input to a gated transformer model. The other insight is to introduce supervised and self-supervised pretraining to improve model performance. We compare TransTab with multiple baseline methods on diverse benchmark datasets and five oncology clinical trial datasets. Overall, TransTab ranks 1.00, 1.00, 1.78 out of 12 methods in supervised learning, incremental feature learning, and transfer learning scenarios, respectively; and the proposed pretraining leads to 2.3% AUC lift on average over the supervised learning.

## [Washing The Unwashable : On The \(Im\)possibility of Fairwashing Detection](#)

- Ali Shahin Shamsabadi · Mohammad Yaghini · Natalie Dullerud · Sierra Wyllie · Ulrich Avodji · Aisha Alaagib · Sébastien Gambs · Nicolas Papernot
- abstract@[open-review](#): The use of black-box models (e.g., deep neural networks) in high-stakes decision-making systems, whose internal logic is complex, raises the need for providing explanations about their decisions. Model explanation techniques mitigate this problem by generating an interpretable and high-fidelity surrogate model (e.g., a logistic regressor or decision tree) to explain the logic of black-box models. In this work, we investigate the issue of fairwashing, in which model explanation techniques are manipulated to rationalize decisions taken by an unfair black-box model using deceptive surrogate models. More precisely, we theoretically characterize and analyze fairwashing, proving that this phenomenon is difficult to avoid due to an irreducible factor—the unfairness of the black-box model. Based on the theory developed, we propose a novel technique, called FRAUD-Detect (Fairness AUDit Detection), to detect fairwashed models by measuring a divergence over subpopulation-wise fidelity measures of the interpretable model. We empirically demonstrate that this divergence is significantly larger in purposefully fairwashed interpretable models than in honest ones. Furthermore, we show that our detector is robust to an informed adversary trying to bypass our detector.

## [Approximate Euclidean lengths and distances beyond Johnson-Lindenstrauss](#)

- Aleksandros Sobczyk · Mathieu Luisier
- abstract@[open-review](#): A classical result of Johnson and Lindenstrauss states that a set of  $n$  high dimensional data points can be projected down to  $\mathcal{O}(\log n/\epsilon^2)$  dimensions such that the square of their pairwise distances is preserved up to a small distortion  $\epsilon \in (0,1)$ . It has been proved that the JL lemma is optimal for the general case, therefore, improvements can only be explored for special cases. This work aims to improve the  $\epsilon^{-2}$  dependency based on techniques inspired by the Hutch++ Algorithm \cite{Meyer2021}, which reduces  $\epsilon^{-2}$  to  $\epsilon^{-1}$  for the related problem of implicit matrix trace estimation. For  $\epsilon=0.01$ , for example, this translates to 100 times less matrix-vector products in the matrix-vector query model to achieve the same accuracy as other previous estimators. We first present an algorithm to estimate the Euclidean lengths of the rows of a matrix. We prove element-wise probabilistic bounds that are at least as good as standard JL approximations in the worst-case, but are asymptotically better for matrices with decaying spectrum. Moreover, for any matrix, regardless its spectrum, the algorithm achieves  $\epsilon$ -accuracy for the total, Frobenius norm-wise relative error using only  $\mathcal{O}(\epsilon^{-1})$  queries. This is a quadratic improvement over the norm-wise error of standard JL approximations. We finally show how these results can be extended to estimate the Euclidean distances between data points and to approximate the statistical leverage scores of a tall-and-skinny data matrix, which are ubiquitous for many applications. Proof-of-concept numerical experiments are presented to validate the theoretical analysis.

## NS3: Neuro-symbolic Semantic Code Search

- Shushan Arakelyan · Anna Hakhverdyan · Miltiadis Allamanis · Christophe Hauser · Luis Garcia · Xiang Ren
- abstract@[open-review](#): Semantic code search is the task of retrieving a code snippet given a textual description of its functionality. Recent work has been focused on using similarity metrics between neural embeddings of text and code. However, current language models are known to struggle with longer, compositional sentences, and multi-step reasoning. To overcome this limitation, we propose supplementing the query sentence with a layout of its semantic structure. The semantic layout is used to break down the final reasoning decision into a series of lower-level decisions. We use a Neural Module Network architecture to implement this idea. We compare our model - \$NS^3\$ (Neuro-Symbolic Semantic Search) - to a number of baselines, including state-of-the-art semantic code retrieval methods, such as CodeBERT, CuBERT and GraphCodeBERT, and evaluate on two datasets - Code Search Net (CSN) and Code Search and Question Answering (CoSQA). On these datasets, we demonstrate that our approach results in higher performance. We also perform additional studies to show the effectiveness of our modular design when handling compositional queries.

## Your Transformer May Not be as Powerful as You Expect

- Shengjie Luo · Shanda Li · Shuxin Zheng · Tie-Yan Liu · Liwei Wang · Di He
- abstract@[open-review](#): Relative Positional Encoding (RPE), which encodes the relative distance between any pair of tokens, is one of the most successful modifications to the original Transformer. As far as we know, theoretical understanding of the RPE-based Transformers is largely unexplored. In this work, we mathematically analyze the power of RPE-based Transformers regarding whether the model is capable of approximating any continuous sequence-to-sequence functions. One may naturally assume the answer is in the affirmative--RPE-based Transformers are universal function approximators. However, we present a negative result by showing there exist continuous sequence-to-sequence functions that RPE-based Transformers cannot approximate no matter how deep and wide the neural network is. One key reason lies in that most RPEs are placed in the softmax attention that always generates a right stochastic matrix. This restricts the network from capturing positional information in the RPEs and limits its capacity. To overcome the problem and make the model more powerful, we first present sufficient conditions for RPE-based Transformers to achieve universal function approximation. With the theoretical guidance, we develop a novel attention module, called Universal RPE-based (URPE) Attention, which satisfies the conditions. Therefore, the corresponding URPE-based Transformers become universal function approximators. Extensive experiments covering typical architectures and tasks demonstrate that our model is parameter-efficient and can achieve superior performance to strong baselines in a wide range of applications.

## Cross-dataset Training Transformers for Robust Action Recognition

- Junwei Liang · Enwei Zhang · Jun Zhang · Chunhua Shen
- abstract@[open-review](#): We study on robust feature representations that can generalize on multiple datasets for action recognition using transformers. Although we have witnessed great progress of action recognition in the past decade, it remains challenging yet valuable how to train a single model that can perform well across multiple datasets. Here we propose a novel multi-dataset training paradigm, MultiTrain, with the design of two new loss terms, namely informative loss and projection loss, aiming to learn robust representations for action recognition. We verify the effectiveness of our method on five challenging datasets, Kinetics-400, Kinetics-700, Moments-in-Time, Activitynet and Something-something-v2 datasets. Extensive experimental results show that our method can consistently improve the state-of-the-art performance. We will release our data, models and code.

## Generalized Delayed Feedback Model with Post-Click Information in Recommender Systems

- Jiaqi Yang · De-Chuan Zhan
- abstract@[open-review](#): Predicting conversion rate (e.g., the probability that a user will purchase an item) is a fundamental problem in machine learning based recommender systems. However, accurate conversion labels are revealed after a long delay, which harms the timeliness of recommender systems. Previous literature concentrates on utilizing early conversions to mitigate such a delayed feedback problem. In this paper, we show that post-click user behaviors are also informative to conversion rate prediction and can be used to improve timeliness. We propose a generalized delayed feedback model (GDFM) that unifies both post-click behaviors and early conversions as stochastic post-click information, which could be utilized to train GDFM in a streaming manner efficiently. Based on GDFM, we further establish a novel perspective that the performance gap introduced by delayed feedback can be attributed to a temporal gap and a sampling gap. Inspired by our analysis, we propose to measure the quality of post-click information with a combination of temporal distance and sample complexity. The training objective is re-weighted accordingly to highlight informative and timely signals. We validate our analysis on public datasets, and experimental performance confirms the effectiveness of our method.

## Skills Regularized Task Decomposition for Multi-task Offline Reinforcement Learning

- Minjong Yoo · SangWoo Cho · Honguk Woo
- abstract@[open-review](#): Reinforcement learning (RL) with diverse offline datasets can have the advantage of leveraging the relation of multiple tasks and the common skills learned across those tasks, hence allowing us to deal with real-world complex problems efficiently in a data-driven way. In offline RL where only offline data is used and online interaction with the environment is restricted, it is yet difficult to achieve the optimal policy for multiple tasks, especially when the data quality varies for the tasks. In this paper, we present a skill-based multi-task RL technique on heterogeneous datasets that are generated by behavior policies of different quality. To learn the shareable knowledge across those datasets effectively, we employ a task decomposition method for which common skills are jointly learned and used as guidance to reformulate a task in shared and achievable subtasks. In this joint learning, we use Wasserstein Auto-Encoder (WAE) to represent both skills and tasks on the same latent space and use the quality-weighted loss as a regularization term to induce tasks to be decomposed into subtasks that are more consistent with high-quality skills than others. To improve the performance of offline RL agents learned on the latent space, we also augment datasets with imaginary trajectories relevant to high-quality skills for each task. Through experiments, we show that our multi-task offline RL approach is robust to different-quality datasets and it outperforms other state-of-the-art algorithms for several robotic manipulation tasks and drone navigation tasks.

## DeepInteraction: Exploring Multi-modal Interaction for 3D Object Detection

- Zeyu Yang · Jiaqi Chen · Zhenwei Miao · Wei Li · Xiatian Zhu · Li Zhang
- abstract@[open-review](#): Existing multi-modal 3D object detectors typically consider a unilateral association strategy with a biased inclination on 3D LiDAR point clouds whilst treating the 2D multi-camera images as an auxiliary information source. As a result, those useful high-resolution information unique with the images is rigidly thrown away in modality association. In essence, the intrinsic complementary nature between the two modalities is fully overlooked by prior arts. In this work, we introduce a novel 3D object detection architecture, dubbed as DeepInteraction, characterized by bilateral interaction and association throughout both representation encoding and decoding, in order to maximally exploit the inter-modal complementary property. Extensive experiments verify the accuracy superiority of DeepInteraction over the state-of-the-art methods by large margin on the large scale nuScenes benchmark.

## Learning Multi-resolution Functional Maps with Spectral Attention for Robust Shape Matching

- Lei Li · Nicolas Donati · Maks Ovsjanikov

- abstract@[open-review](#): In this work, we present a novel non-rigid shape matching framework based on multi-resolution functional maps with spectral attention. Indeed, existing functional map learning methods all rely on a choice of the critical spectral resolution hyper-parameter, which can severely affect the overall accuracy or lead to overfitting, if not chosen carefully. In this paper, we show that spectral resolution tuning can be alleviated by introducing spectral attention. Our framework is applicable in both supervised and unsupervised settings, and we show that it is possible to train the network so that it can adapt the spectral resolution, depending on the given shape input. More specifically, we propose to compute multi-resolution functional maps that characterize correspondence across a wide range of spectral resolution, and introduce a spectral attention network that helps to combine this representation into a single coherent final correspondence. Our approach is not only accurate with near-isometric input, for which a high spectral resolution is typically preferred, but also robust and able to produce reasonable matching even in the presence of significant distortion, which poses great challenges to existing methods. We demonstrate the superior performance of our approach through experiments on a suite of challenging non-rigid shape matching benchmarks, including a new non-isometric correspondence dataset.

## [The Unreasonable Effectiveness of Fully-Connected Layers for Low-Data Regimes](#)

- Peter Kocsis · Ismail Elezi · Peter Säken · Guillem Braso · Matthias Niessner · Laura Leal-Taix ©
- abstract@[open-review](#): Convolutional neural networks were the standard for solving many computer vision tasks until recently, when Transformers of MLP-based architectures have started to show competitive performance. These architectures typically have a vast number of weights and need to be trained on massive datasets; hence, they are not suitable for their use in low-data regimes. In this work, we propose a simple yet effective framework to improve generalization from small amounts of data. We augment modern CNNs with fully-connected (FC) layers and show the massive impact this architectural change has in low-data regimes. We further present an online joint knowledge-distillation method to utilize the extra FC layers at train time but avoid them during test time. This allows us to improve the generalization of a CNN-based model without any increase in the number of weights at test time. We perform classification experiments for a large range of network backbones and several standard datasets on supervised learning and active learning. Our experiments significantly outperform the networks without fully-connected layers, reaching a relative improvement of up to 16% validation accuracy in the supervised setting without adding any extra parameters during inference.

## [FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness](#)

- Tri Dao · Dan Fu · Stefano Ermon · Atri Rudra · Christopher R ©
- abstract@[open-review](#): Transformers are slow and memory-hungry on long sequences, since the time and memory complexity of self-attention are quadratic in sequence length. Approximate attention methods have attempted to address this problem by trading off model quality to reduce the compute complexity, but often do not achieve wall-clock speedup. We argue that a missing principle is making attention algorithms IO-aware---accounting for reads and writes between levels of GPU memory. We propose FlashAttention, an IO-aware exact attention algorithm that uses tiling to reduce the number of memory reads/writes between GPU high bandwidth memory (HBM) and GPU on-chip SRAM. We analyze the IO complexity of FlashAttention, showing that it requires fewer HBM accesses than standard attention, and is optimal for a range of SRAM sizes. We also extend FlashAttention, yielding an approximate attention algorithm that is faster than any existing approximate attention method. FlashAttention, 3x speedup on GPT-2 (seq. length 1K), and 2.4x speedup on long-range arena (seq. length 1K-4K). FlashAttention, yielding higher quality models (0.7 better perplexity on GPT-2 and 6.4 points of lift on long-document classification) and entirely new capabilities: the first Transformers to achieve better-than-chance performance on the Path-X challenge (seq. length 16K, 61.4% accuracy) and Path-256 (seq. length 64K, 63.1% accuracy).

## [GAPX: Generalized Autoregressive Paraphrase-Identification X](#)

- Yifei Zhou · Renyu Li · Hayden Housen · Ser Nam Lim
- abstract@[open-review](#): Paraphrase Identification is a fundamental task in Natural Language Processing. While much progress has been made in the field, the performance of many state-of-the-art models often suffer from distribution shift during inference time. We verify that a major source of this performance drop comes from biases introduced by negative examples. To overcome these biases, we propose in this paper to train two separate models, one that only utilizes the positive pairs and the other the negative pairs. This enables us the option of deciding how much to utilize the negative model, for which we introduce a perplexity based out-of-distribution metric that we show can effectively and automatically determine how much weight it should be given during inference. We support our findings with strong empirical results.

## [Geo-SIC: Learning Deformable Geometric Shapes in Deep Image Classifiers](#)

- Jian Wang · Miaomiao Zhang
- abstract@[open-review](#): Deformable shapes provide important and complex geometric features of objects presented in images. However, such information is oftentimes missing or underutilized as implicit knowledge in many image analysis tasks. This paper presents Geo-SIC, the first deep learning model to learn deformable shapes in a deformation space for an improved performance of image classification. We introduce a newly designed framework that (i) simultaneously derives features from both image and latent shape spaces with large intra-class variations; and (ii) gains increased model interpretability by allowing direct access to the underlying geometric features of image data. In particular, we develop a boosted classification network, equipped with an unsupervised learning of geometric shape representations characterized by diffeomorphic transformations within each class. In contrast to previous approaches using pre-extracted shapes, our model provides a more fundamental approach by naturally learning the most relevant shape features jointly with an image classifier. We demonstrate the effectiveness of our method on both simulated 2D images and real 3D brain magnetic resonance (MR) images. Experimental results show that our model substantially improves the image classification accuracy with an additional benefit of increased model interpretability. All code and data produced in this research will be made publicly available online.

## [Learning General World Models in a Handful of Reward-Free Deployments](#)

- Jack Parker-Holder · Yingchen Xu · Philip Ball · Aldo Pacchiano · Oleh Rybkin · S Roberts · Tim Rocktäschel · Edward Grefenstette
- abstract@[open-review](#): Building generally capable agents is a grand challenge for deep reinforcement learning (RL). To approach this challenge practically, we outline two key desiderata: 1) to facilitate generalization, exploration should be task agnostic; 2) to facilitate scalability, exploration policies should collect large quantities of data without costly centralized retraining. Combining these two properties, we introduce the reward-free deployment efficiency setting, a new paradigm for RL research. We then present CASCADE, a novel approach for self-supervised exploration in this new setting. CASCADE seeks to learn a world model by collecting data with a population of agents, using an information theoretic objective inspired by Bayesian Active Learning. CASCADE achieves this by specifically maximizing the diversity of trajectories sampled by the population through a novel cascading objective. We show a tabular version of CASCADE theoretically improves upon naïve approaches that do not account for population diversity. We then demonstrate that CASCADE collects diverse task-agnostic datasets and learns agents that generalize zero-shot to novel, unseen downstream tasks on Atari, MiniGrid and the DM Control Suite.

## [Explainable Reinforcement Learning via Model Transforms](#)

- Mira Finkelstein · Nitsan Levy · Lucy Liu · Yoav Kolumbus · David Parkes · Jeffrey S Rosenschein · Sarah Keren
- abstract@[open-review](#): Understanding emerging behaviors of reinforcement learning (RL) agents may be difficult since such agents are often trained in complex environments using highly complex decision making procedures. This has given rise to a variety of approaches to explainability in RL that aim to reconcile discrepancies that may arise between the behavior of an agent and the behavior that is anticipated by an observer. Most recent approaches have relied either on domain knowledge, that may not always be available, on an analysis of the agent's policy, or on an analysis of specific elements of the

underlying environment, typically modeled as a Markov Decision Process (MDP). Our key claim is that even if the underlying MDP is not fully known (e.g., the transition probabilities have not been accurately learned) or is not maintained by the agent (i.e., when using model-free methods), it can nevertheless be exploited to automatically generate explanations. For this purpose, we suggest using formal MDP abstractions and transforms, previously used in the literature for expediting the search for optimal policies, to automatically produce explanations. Since such transforms are typically based on a symbolic representation of the environment, they may represent meaningful explanations for gaps between the anticipated and actual agent behavior. We formally define this problem, suggest a class of transforms that can be used for explaining emergent behaviors, and suggest methods that enable efficient search for an explanation. We demonstrate the approach on a set of standard benchmarks.

## [Markov Chain Score Ascent: A Unifying Framework of Variational Inference with Markovian Gradients](#)

- Kyurae Kim · Jisu Oh · Jacob Gardner · Adji Bousso Dieng · Hongseok Kim
- abstract@[open-review](#): Minimizing the inclusive Kullback-Leibler (KL) divergence with stochastic gradient descent (SGD) is challenging since its gradient is defined as an integral over the posterior. Recently, multiple methods have been proposed to run SGD with biased gradient estimates obtained from a Markov chain. This paper provides the first non-asymptotic convergence analysis of these methods by establishing their mixing rate and gradient variance. To do this, we demonstrate that these methods—“which we collectively refer to as Markov chain score ascent (MCSA) methods”—can be cast as special cases of the Markov chain gradient descent framework. Furthermore, by leveraging this new understanding, we develop a novel MCSA scheme, parallel MCSA (pMCSA), that achieves a tighter bound on the gradient variance. We demonstrate that this improved theoretical result translates to superior empirical performance.

## [Local Latent Space Bayesian Optimization over Structured Inputs](#)

- Natalie Maus · Haydn Jones · Juston Moore · Matt Kusner · John Bradshaw · Jacob Gardner
- abstract@[open-review](#): Bayesian optimization over the latent spaces of deep autoencoder models (DAEs) has recently emerged as a promising new approach for optimizing challenging black-box functions over structured, discrete, hard-to-enumerate search spaces (e.g., molecules). Here the DAE dramatically simplifies the search space by mapping inputs into a continuous latent space where familiar Bayesian optimization tools can be more readily applied. Despite this simplification, the latent space typically remains high-dimensional. Thus, even with a well-suited latent space, these approaches do not necessarily provide a complete solution, but may rather shift the structured optimization problem to a high-dimensional one. In this paper, we propose LOL-BO, which adapts the notion of trust regions explored in recent work on high-dimensional Bayesian optimization to the structured setting. By reformulating the encoder to function as both an encoder for the DAE globally and as a deep kernel for the surrogate model within a trust region, we better align the notion of local optimization in the latent space with local optimization in the input space. LOL-BO achieves as much as 20 times improvement over state-of-the-art latent space Bayesian optimization methods across six real-world benchmarks, demonstrating that improvement in optimization strategies is as important as developing better DAE models.

## [A Hybrid Neural Autoencoder for Sensory Neuroprostheses and Its Applications in Bionic Vision](#)

- Jacob Granley · Lucas Relic · Michael Beyeler
- abstract@[open-review](#): Sensory neuroprostheses are emerging as a promising technology to restore lost sensory function or augment human capacities. However, sensations elicited by current devices often appear artificial and distorted. Although current models can predict the neural or perceptual response to an electrical stimulus, an optimal stimulation strategy solves the inverse problem: what is the required stimulus to produce a desired response? Here we frame this as an end-to-end optimization problem, where a deep neural network encoder is trained to invert a known, fixed forward model that approximates the underlying biological system. As a proof of concept, we demonstrate the effectiveness of our hybrid neural autoencoder (HNA) on the use case of visual neuroprostheses. We found that HNA is able to produce high-fidelity stimuli from the MNIST and COCO datasets that outperform conventional encoding strategies and surrogate techniques across all tested conditions.

## [Unsupervised Domain Adaptation for Semantic Segmentation using Depth Distribution](#)

- Quanliang Wu · Huajun Liu
- abstract@[open-review](#): Recent years have witnessed significant advancements made in the field of unsupervised domain adaptation for semantic segmentation. Depth information has been proved to be effective in building a bridge between synthetic datasets and real-world datasets. However, the existing methods may not pay enough attention to depth distribution in different categories, which makes it possible to use them for further improvement. Besides the existing methods that only use depth regression as an auxiliary task, we propose to use depth distribution density to support semantic segmentation. Therefore, considering the relationship among depth distribution density, depth and semantic segmentation, we also put forward a branch balance loss for these three subtasks in multi-task learning schemes. In addition, we also propose a spatial aggregation priors of pixels in different categories, which is used to refine the pseudo-labels for self-training, thus further improving the performance of the prediction model. Experiments on SYNTHIA-to-Cityscapes and SYNTHIA-to-Mapillary benchmarks show the effectiveness of our proposed method.

## [Visual correspondence-based explanations improve AI robustness and human-AI team accuracy](#)

- Mohammad Reza Taesiri · Giang Nguyen · Anh Nguyen
- abstract@[open-review](#): Explaining artificial intelligence (AI) predictions is increasingly important and even imperative in many high-stake applications where humans are the ultimate decision-makers. In this work, we propose two novel architectures of explainable image classifiers that first explain, and then predict (as opposed to post-hoc explanation methods). Our models first rank the training-set images by their distance with the query in an image-level deep feature space. And then, we re-rank the top-50 shortlisted candidates using patch-wise similarity of 5 highest-similarity pairs of patches between the query and every candidate. On ImageNet, our models improve (by 1-4 points) the out-of-distribution accuracy on several datasets including Adversarial Patch and ImageNet-R while performing marginally worse (by 1-2 points) on ImageNet to the baselines (ResNet-50 pre-trained ImageNet). A consistent trend is observed on CUB. Via a large-scale, human study (~60 users per method per dataset) on ImageNet and CUB, we find our proposed correspondence-based explanations led to human-alone image classification accuracy and human-AI team accuracy that are consistently better than those of k-NN. Our correspondence-based explanations help users better correctly reject AI's wrong decisions than all other tested methods. Interestingly, for the first time, we show that it is possible to achieve complementary human-AI team accuracy (i.e. that is higher than either AI-alone or human-alone), in both image classification tasks.

## [Interaction Modeling with Multiplex Attention](#)

- Fan-Yun Sun · Isaac Kauvar · Ruohan Zhang · Jiachen Li · Mykel J Kochenderfer · Jiajun Wu · Nick Haber
- abstract@[open-review](#): Modeling multi-agent systems requires understanding how agents interact. Such systems are often difficult to model because they can involve a variety of types of interactions that layer together to drive rich social behavioral dynamics. Here we introduce a method for accurately modeling multi-agent systems. We present Interaction Modeling with Multiplex Attention (IMMA), a forward prediction model that uses a multiplex latent graph to represent multiple independent types of interactions and attention to account for relations of different strengths. We also introduce Progressive Layer Training, a training strategy for this architecture. We show that our approach outperforms state-of-the-art models in trajectory forecasting and relation inference, spanning three multi-agent scenarios: social navigation, cooperative task achievement, and team sports. We further demonstrate that our approach can improve zero-shot generalization and allows us to probe how different interactions impact agent behavior.

## Diffusion Curvature for Estimating Local Curvature in High Dimensional Data

- Dhananjay Bhaskar · Kincaid MacDonald · Oluwadamilola Fasina · Dawson Thomas · Bastian Rieck · Ian Adelstein · Smita Krishnaswamy
- abstract@[open-review](#): We introduce a new intrinsic measure of local curvature on point-cloud data called diffusion curvature. Our measure uses the framework of diffusion maps, including the data diffusion operator, to structure point cloud data and define local curvature based on the laziness of a random walk starting at a point or region of the data. We show that this laziness directly relates to volume comparison results from Riemannian geometry. We then extend this scalar curvature notion to an entire quadratic form using neural network estimations based on the diffusion map of point-cloud data. We show applications of both estimations on toy data, single-cell data, and on estimating local Hessian matrices of neural network loss landscapes.

## FedSR: A Simple and Effective Domain Generalization Method for Federated Learning

- A. Tuan Nguyen · Ser Nam Lim · Philip Torr
- abstract@[open-review](#): Federated Learning (FL) refers to the decentralized and privacy-preserving machine learning framework in which multiple clients collaborate (with the help of a central server) to train a global model without sharing their data. However, most existing FL methods only focus on maximizing the model's performance on the source clients' data (e.g., mobile users) without considering its generalization ability to unknown target data (e.g., a new user). In this paper, we incorporate the problem of Domain Generalization (DG) into Federated Learning to tackle the aforementioned issue. However, virtually all existing DG methods require a centralized setting where data is shared across the domains, which violates the principles of decentralized FL and hence not applicable. To this end, we propose a simple yet novel representation learning framework, namely FedSR, which enables domain generalization while still respecting the decentralized and privacy-preserving natures of this FL setting. Motivated by classical machine learning algorithms, we aim to learn a simple representation of the data for better generalization. In particular, we enforce an L2-norm regularizer on the representation and a conditional mutual information (between the representation and the data given the label) regularizer to encourage the model to only learn essential information (while ignoring spurious correlations such as the background). Furthermore, we provide theoretical connections between the above two objectives and representation alignment in domain generalization. Extensive experimental results suggest that our method significantly outperforms relevant baselines in this particular problem.

## Unifying Information Extraction with Latent Adaptive Structure-aware Generative Language Model

- Hao Fei · Shengqiong Wu · Libo Qin · Jingye Li · Bobo Li · Fei Li · Meishan Zhang · Min Zhang · Tat-Seng Chua
- abstract@[open-review](#): Universally modeling all typical information extraction tasks (UIE) with one generative language model (GLM) has revealed great potential by the latest study, where various IE predictions are unified into a linearized hierarchical expression under a GLM. Syntactic structure information, a type of effective feature which has been extensively utilized in IE community, should also be beneficial to UIE. In this work, we propose a novel structure-aware GLM, fully unleashing the power of syntactic knowledge for UIE. A heterogeneous structure inductor is explored to unsupervisedly induce rich heterogeneous structural representations by post-training an existing GLM. In particular, a structural broadcaster is devised to compact various latent trees into explicit high-order forests, helping to guide a better generation during decoding. We finally introduce a task-oriented structure fine-tuning mechanism, further adjusting the learned structures to most coincide with the end-task's need. Over 12 IE benchmarks across 7 tasks our system shows significant improvements over the baseline UIE system. Further in-depth analyses show that our GLM learns rich task-adaptive structural bias that greatly resolves the UIE crux, the long-range dependence issue and boundary identifying.

## Physically-Based Face Rendering for NIR-VIS Face Recognition

- Yunqi Miao · Alexandros Lattas · Jiankang Deng · Jungong Han · Stefanos Zafeiriou
- abstract@[open-review](#): Near infrared (NIR) to Visible (VIS) face matching is challenging due to the significant domain gaps as well as a lack of sufficient data for cross-modality model training. To overcome this problem, we propose a novel method for paired NIR-VIS facial images generation. Specifically, we reconstruct 3D face shape and reflectance from a large 2D facial dataset and introduce a novel method of transforming the VIS reflectance to NIR reflectance. We then use a physically-based renderer to generate a vast, high-resolution and photorealistic dataset consisting of various poses and identities in the NIR and VIS spectra. Moreover, to facilitate the identity feature learning, we propose an IDentity-based Maximum Mean Discrepancy (ID-MMD) loss, which not only reduces the modality gap between NIR and VIS images at the domain level but encourages the network to focus on the identity features instead of facial details, such as poses and accessories. Extensive experiments conducted on four challenging NIR-VIS face recognition benchmarks demonstrate that the proposed method can achieve comparable performance with the state-of-the-art (SOTA) methods without requiring any existing NIR-VIS face recognition datasets. With slightly fine-tuning on the target NIR-VIS face recognition datasets, our method can significantly surpass the SOTA performance.

## Unsupervised Learning From Incomplete Measurements for Inverse Problems

- Julián Tachella · Dongdong Chen · Mike Davies
- abstract@[open-review](#): In many real-world inverse problems, only incomplete measurement data are available for training which can pose a problem for learning a reconstruction function. Indeed, unsupervised learning using a fixed incomplete measurement process is impossible in general, as there is no information in the nullspace of the measurement operator. This limitation can be overcome by using measurements from multiple operators. While this idea has been successfully applied in various applications, a precise characterization of the conditions for learning is still lacking. In this paper, we fill this gap by presenting necessary and sufficient conditions for learning the underlying signal model needed for reconstruction which indicate the interplay between the number of distinct measurement operators, the number of measurements per operator, the dimension of the model and the dimension of the signals. Furthermore, we propose a novel and conceptually simple unsupervised learning loss which only requires access to incomplete measurement data and achieves a performance on par with supervised learning when the sufficient condition is verified. We validate our theoretical bounds and demonstrate the advantages of the proposed unsupervised loss compared to previous methods via a series of experiments on various imaging inverse problems, such as accelerated magnetic resonance imaging, compressed sensing and image inpainting.

## Dynamic 3D from Monocular Video: Reality Check

- Hang Gao · Ruilong Li · Shubham Tulsiani · Bryan Russell · Angjoo Kanazawa
- abstract@[open-review](#): We study the recent progress on inferring dynamic 3D scene representations from a monocular video sequence. We take a closer look and find that although recent approaches demonstrate impressive results, they are evaluated on datasets with a protocol that effectively provide multi-view signals. We propose effective multi-view factors (EMF) to quantify the amount of multi-view signal in a sequence based on the magnitude of relative camera and scene motion. We then propose a new dataset of monocular videos with depth data and evaluation protocols based on depth-guided masked PSNR and correspondence. We evaluate representative approaches on this dataset and find that many challenges remain in the monocular setup. We advise future works to report EMF on their input sequences in order to assess the difficulty of the task. We will release our dataset and code including the metric and the dataset generation in hopes that it will be a useful toolbox for future work.

## Look More but Care Less in Video Recognition

- Yitian Zhang · Yue Bai · Huan Wang · Yi Xu · Yun Fu

- abstract@[open-review](#): Existing action recognition methods typically sample a few frames to represent each video to avoid the enormous computation, which often limits the recognition performance. To tackle this problem, we propose Ample and Focal Network (AFNet), which is composed of two branches to utilize more frames but with less computation. Specifically, the Ample Branch takes all input frames to obtain abundant information with condensed computation and provides the guidance for Focal Branch by the proposed Navigation Module; the Focal Branch squeezes the temporal size to only focus on the salient frames at each convolution block; in the end, the results of two branches are adaptively fused to prevent the loss of information. With this design, we can introduce more frames to the network but cost less computation. Besides, we demonstrate AFNet can utilize less frames while achieving higher accuracy as the dynamic selection in intermediate features enforces implicit temporal modeling. Further, we show that our method can be extended to reduce spatial redundancy with even less cost. Extensive experiments on five datasets demonstrate the effectiveness and efficiency of our method.

## [Peripheral Vision Transformer](#)

- Juhong Min · Yucheng Zhao · Chong Luo · Minsu Cho
- abstract@[open-review](#): Human vision possesses a special type of visual processing systems called peripheral vision. Partitioning the entire visual field into multiple contour regions based on the distance to the center of our gaze, the peripheral vision provides us the ability to perceive various visual features at different regions. In this work, we take a biologically inspired approach and explore to model peripheral vision in deep neural networks for visual recognition. We propose to incorporate peripheral position encoding to the multi-head self-attention layers to let the network learn to partition the visual field into diverse peripheral regions given training data. We evaluate the proposed network, dubbed PerViT, on ImageNet-1K and systematically investigate the inner workings of the model for machine perception, showing that the network learns to perceive visual data similarly to the way that human vision does. The performance improvements in image classification over the baselines across different model sizes demonstrate the efficacy of the proposed method.

## [On the Global Convergence Rates of Decentralized Softmax Gradient Play in Markov Potential Games](#)

- Runyu Zhang · Jincheng Mei · Bo Dai · Dale Schuurmans · Na Li
- abstract@[open-review](#): Softmax policy gradient is a popular algorithm for policy optimization in single-agent reinforcement learning, particularly since projection is not needed for each gradient update. However, in multi-agent systems, the lack of central coordination introduces significant additional difficulties in the convergence analysis. Even for a stochastic game with identical interest, there can be multiple Nash Equilibria (NEs), which disables proof techniques that rely on the existence of a unique global optimum. Moreover, the softmax parameterization introduces non-NE policies with zero gradient, making it difficult for gradient-based algorithms in seeking NEs. In this paper, we study the finite time convergence of decentralized softmax gradient play in a special form of game, Markov Potential Games (MPGs), which includes the identical interest game as a special case. We investigate both gradient play and natural gradient play, with and without  $\log$ -barrier regularization. The established convergence rates for the unregularized cases contain a trajectory dependent constant that can be arbitrarily large, whereas the  $\log$ -barrier regularization overcomes this drawback, with the cost of slightly worse dependence on other factors such as the action set size. An empirical study on an identical interest matrix game confirms the theoretical findings.

## [On Uncertainty, Tempering, and Data Augmentation in Bayesian Classification](#)

- Sanyam Kapoor · Wesley Maddox · Pavel Izmailov · Andrew Wilson
- abstract@[open-review](#): Aleatoric uncertainty captures the inherent randomness of the data, such as measurement noise. In Bayesian regression, we often use a Gaussian observation model, where we control the level of aleatoric uncertainty with a noise variance parameter. By contrast, for Bayesian classification we use a categorical distribution with no mechanism to represent our beliefs about aleatoric uncertainty. Our work shows that explicitly accounting for aleatoric uncertainty significantly improves the performance of Bayesian neural networks. We note that many standard benchmarks, such as CIFAR-10, have essentially no aleatoric uncertainty. Moreover, we show data augmentation in approximate inference softens the likelihood, leading to underconfidence and profoundly misrepresenting our honest beliefs about aleatoric uncertainty. Accordingly, we find that a cold posterior, tempered by a power greater than one, often more honestly reflects our beliefs about aleatoric uncertainty than no tempering --- providing an explicit link between data augmentation and cold posteriors. We show that we can match or exceed the performance of posterior tempering by using a Dirichlet observation model, where we explicitly control the level of aleatoric uncertainty, without any need for tempering.

## [Simple Mechanisms for Welfare Maximization in Rich Advertising Auctions](#)

- Gagan Aggarwal · Kshipra Bhawalkar · Aranyak Mehta · Divyarthi Mohan · Alexandros Psomas
- abstract@[open-review](#): Internet ad auctions have evolved from a few lines of text to richer informational layouts that include images, sitelinks, videos, etc. Ads in these new formats occupy varying amounts of space, and an advertiser can provide multiple formats, only one of which can be shown. The seller is now faced with a multi-parameter mechanism design problem. Computing an efficient allocation is computationally intractable, and therefore the standard Vickrey-Clarke-Groves (VCG) auction, while truthful and welfare-optimal, is impractical. In this paper, we tackle a fundamental problem in the design of modern ad auctions. We adopt a Myersonian'' approach and study allocation rules that are monotone both in the bid and set of rich ads. We show that such rules can be paired with a payment function to give a truthful auction. Our main technical challenge is designing a monotone rule that yields a good approximation to the optimal welfare. Monotonicity doesn't hold for standard algorithms, e.g. the incremental bang-per-buck order, that give good approximations to knapsack-like" problems such as ours. In fact, we show that no deterministic monotone rule can approximate the optimal welfare within a factor better than  $2\sqrt{2}$  (while there is a non-monotone FPTAS). Our main result is a new, simple, greedy and monotone allocation rule that guarantees a  $3\sqrt{3}$  approximation. In ad auctions in practice, monotone allocation rules are often paired with the so-called Generalized Second Price (GSP) payment rule, which charges the minimum threshold price below which the allocation changes. We prove that, even though our monotone allocation rule paired with GSP is not truthful, its Price of Anarchy (PoA) is bounded. Under standard no-overbidding assumptions, we prove bounds on the a pure and Bayes-Nash PoA. Finally, we experimentally test our algorithms on real-world data.

## [Are all Frames Equal? Active Sparse Labeling for Video Action Detection](#)

- Aayush Rana · Yogesh Rawat
- abstract@[open-review](#): Video action detection requires annotations at every frame, which drastically increases the labeling cost. In this work, we focus on efficient labeling of videos for action detection to minimize this cost. We propose active sparse labeling (ASL), a novel active learning strategy for video action detection. Sparse labeling will reduce the annotation cost but poses two main challenges; 1) how to estimate the utility of annotating a single frame for action detection as detection is performed at video level?, and 2) how these sparse labels can be used for action detection which require annotations on all the frames? This work attempts to address these challenges within a simple active learning framework. For the first challenge, we propose a novel frame-level scoring mechanism aimed at selecting most informative frames in a video. Next, we introduce a novel loss formulation which enables training of action detection model with these sparsely selected frames. We evaluate the proposed approach on two different action detection benchmark datasets, UCF-101-24 and J-HMDB-21, and observed that active sparse labeling can be very effective in saving annotation costs. We demonstrate that the proposed approach performs better than random selection, outperforming all other baselines, with performance comparable to supervised approach using merely 10% annotations.

## [Influencing Long-Term Behavior in Multiagent Reinforcement Learning](#)

- Dong-Ki Kim · Matthew Riemer · Miao Liu · Jakob Foerster · Michael Everett · Chuangchuang Sun · Gerald Tesauro · Jonathan How
- abstract@[open-review](#): The main challenge of multiagent reinforcement learning is the difficulty of learning useful policies in the presence of other simultaneously learning agents whose changing behaviors jointly affect the environment's transition and reward dynamics. An effective approach that has recently emerged for addressing this non-stationarity is for each agent to anticipate the learning of other agents and influence the evolution of future policies towards desirable behavior for its own benefit. Unfortunately, previous approaches for achieving this suffer from myopic evaluation, considering only a finite number of policy updates. As such, these methods can only influence transient future policies rather than achieving the promise of scalable equilibrium selection approaches that influence the behavior at convergence. In this paper, we propose a principled framework for considering the limiting policies of other agents as time approaches infinity. Specifically, we develop a new optimization objective that maximizes each agent's average reward by directly accounting for the impact of its behavior on the limiting set of policies that other agents will converge to. Our paper characterizes desirable solution concepts within this problem setting and provides practical approaches for optimizing over possible outcomes. As a result of our farsighted objective, we demonstrate better long-term performance than state-of-the-art baselines across a suite of diverse multiagent benchmark domains.

## [A Practical, Progressively-Expressive GNN](#)

- Lingxiao Zhao · Neil Shah · Leman Akoglu
- abstract@[open-review](#): Message passing neural networks (MPNNs) have become a dominant flavor of graph neural networks (GNNs) in recent years. Yet, MPNNs come with notable limitations; namely, they are at most as powerful as the 1-dimensional Weisfeiler-Leman (1-WL) test in distinguishing graphs in a graph isomorphism testing frame-work. To this end, researchers have drawn inspiration from the k-WL hierarchy to develop more expressive GNNs. However, current k-WL-equivalent GNNs are not practical for even small values of k, as k-WL becomes combinatorially more complex as k grows. At the same time, several works have found great empirical success in graph learning tasks without highly expressive models, implying that chasing expressiveness with a  $\infty$ -coarse-grained ruler of expressivity like k-WL is often unneeded in practical tasks. To truly understand the expressiveness-complexity tradeoff, one desires a more  $\infty$ -fine-grained ruler, which can more gradually increase expressiveness. Our work puts forth such a proposal: Namely, we first propose the  $(k, c)$ -SETWL hierarchy with greatly reduced complexity from k-WL, achieved by moving from k-tuples of nodes to sets with  $k$  nodes defined over  $c$  connected components in the induced original graph. We show favorable theoretical results for this model in relation to k-WL, and concretize it via  $(k, c)$ -SETGNN, which is as expressive as  $(k, c)$ -SETWL. Our model is practical and progressively-expressive, increasing in power with k and c. We demonstrate effectiveness on several benchmark datasets, achieving several state-of-the-art results with runtime and memory usage applicable to practical graphs.

## [Constrained Predictive Coding as a Biologically Plausible Model of the Cortical Hierarchy](#)

- Siavash Golkar · Tiberiu Tesileanu · Yanis Bahroun · Anirvan Sengupta · Dmitri Chklovskii
- abstract@[open-review](#): Predictive coding (PC) has emerged as an influential normative model of neural computation, with numerous extensions and applications. As such, much effort has been put into mapping PC faithfully onto the cortex, but there are issues that remain unresolved or controversial. In particular, current implementations often involve separate value and error neurons and require symmetric forward and backward weights across different brain regions. These features have not been experimentally confirmed. In this work, we show that the PC framework in the linear regime can be modified to map faithfully onto the cortical hierarchy in a manner compatible with empirical observations. By employing a disentangling-inspired constraint on hidden-layer neural activities, we derive an upper bound for the PC objective. Optimization of this upper bound leads to an algorithm that shows the same performance as the original objective and maps onto a biologically plausible network. The units of this network can be interpreted as multi-compartmental neurons with non-Hebbian learning rules, with a remarkable resemblance to recent experimental findings. There exist prior models which also capture these features, but they are phenomenological, while our work is a normative derivation. Notably, the network we derive does not involve one-to-one connectivity or signal multiplexing, which the phenomenological models required, indicating that these features are not necessary for learning in the cortex. The normative nature of our algorithm in the simplified linear case allows us to prove interesting properties of the framework and analytically understand the computational role of our network's components. The parameters of our network have natural interpretations as physiological quantities in a multi-compartmental model of pyramidal neurons, providing a concrete link between PC and experimental measurements carried out in the cortex.

## [SPDNet: A Large-Scale Imagery Dataset and Benchmark for Spatial Precipitation Downscaling](#)

- Xuanhong Chen · Kairui Feng · Bingbing Ni · Naiyuan Liu · Yifan Lu · Ziang Liu · Zhengyan Tong
- abstract@[open-review](#): AI-for-science approaches have been applied to solve scientific problems (e.g., nuclear fusion, ecology, genomics, meteorology) and have achieved highly promising results. Spatial precipitation downscaling is one of the most important meteorological issues and urgently requires the participation of AI. However, the lack of a well-organized and annotated large-scale dataset hinders the training and verification of more effective and advancing deep-learning models for precipitation downscaling. To alleviate these obstacles, we present the first large-scale spatial precipitation downscaling dataset named SPDNet, which contains more than \$62,400\$ pairs of high-quality low/high-resolution precipitation maps for over \$17\$ years, ready to help the evolution of deep learning models in precipitation downscaling. Specifically, the precipitation maps carefully collected in SPDNet cover various meteorological phenomena (e.g., hurricane, squall), which is of great help to improve the model generalization ability. In addition, the map pairs in SPDNet are organized in the form of image sequences (\$720\$ maps per month or \$1\$ map/hour), showing complex physical properties, temporal misalignment, temporal sparse, and fluid properties. Two deep-learning-oriented metrics are specifically introduced to evaluate or verify the comprehensive performance of the trained model, (e.g., prediction maps reconstruction accuracy). To illustrate the applications of SPDNet, \$14\$ state-of-the-art models, including deep models and traditional approaches, are evaluated. To fully explore potential downscaling solutions, we propose an implicit physical estimation benchmark framework to learn the above characteristics. Extensive experiments demonstrate the value of SPDNet in training and evaluating downscaling models.

## [Semi-supervised Vision Transformers at Scale](#)

- Zhaowei Cai · Avinash Ravichandran · Paolo Favaro · Manchen Wang · Davide Modolo · Rahul Bhotika · Zhuowen Tu · Stefano Soatto
- abstract@[open-review](#): We study semi-supervised learning (SSL) for vision transformers (ViT), an under-explored topic despite the wide adoption of the ViT architectures to different tasks. To tackle this problem, we propose a new SSL pipeline, consisting of first un/self-supervised pre-training, followed by supervised fine-tuning, and finally semi-supervised fine-tuning. At the semi-supervised fine-tuning stage, we adopt an exponential moving average (EMA)-Teacher framework instead of the popular FixMatch, since the former is more stable and delivers higher accuracy for semi-supervised vision transformers. In addition, we propose a probabilistic pseudo mixup mechanism to interpolate unlabeled samples and their pseudo labels for improved regularization, which is important for training ViTs with weak inductive bias. Our proposed method, dubbed Semi-ViT, achieves comparable or better performance than the CNN counterparts in the semi-supervised classification setting. Semi-ViT also enjoys the scalability benefits of ViTs that can be readily scaled up to large-size models with increasing accuracies. For example, Semi-ViT-Huge achieves an impressive \$80\%\$ top-1 accuracy on ImageNet using only \$1\%\$ labels, which is comparable with Inception-v4 using \$100\%\$ ImageNet labels.

## [Deep Fourier Up-Sampling](#)

- man zhou · Hu Yu · Jie Huang · Feng Zhao · Jinwei Gu · Chen Change Loy · Deyu Meng · Chongyi Li
- abstract@[open-review](#): Existing convolutional neural networks widely adopt spatial down-/up-sampling for multi-scale modeling. However, spatial up-sampling operators (e.g., interpolation, transposed convolution, and un-pooling) heavily depend on local pixel attention, incapably exploring the global dependency. In contrast, the Fourier domain is in accordance with the nature of global modeling according to the spectral convolution theorem. Unlike the spatial domain that easily performs up-sampling with the property of local similarity, up-sampling in the Fourier domain is more challenging as it does not

follow such a local property. In this study, we propose a theoretically feasible Deep Fourier Up-Sampling (FourierUp) to solve these issues. We revisit the relationships between spatial and Fourier domains and reveal the transform rules on the features of different resolutions in the Fourier domain, which provide key insights for FourierUp's designs. FourierUp as a generic operator consists of three key components: 2D discrete Fourier transform, Fourier dimension increase rules, and 2D inverse Fourier transform, which can be directly integrated with existing networks. Extensive experiments across multiple computer vision tasks, including object detection, image segmentation, image de-raining, image dehazing, and guided image super-resolution, demonstrate the consistent performance gains obtained by introducing our FourierUp. Code will be publicly available.

## [Free probability as a solution to the problem of tuning neural networks](#)

- Reda CHHAIBI · Tariq Daouda · Ezechiel Kahn
- abstract@[open-review](#): Gradient descent during the learning process of a neural network can be subject to many instabilities. The spectral density of the Jacobian is a key component for analyzing stability. Following the works of Pennington et al., such Jacobians are modeled using free multiplicative convolutions from Free Probability Theory (FPT). We present a reliable and very fast method for computing the associated spectral densities, for given architecture and initialization. This method has a controlled and proven convergence. Our technique is based on an homotopy method: it is an adaptative Newton-Raphson scheme which chains basins of attraction. We find contiguous lily pad-like basins and step from one to the next, heading towards the objective. In order to demonstrate the relevance of our method we show that the relevant FPT metrics computed before training are highly correlated to final test losses up to 85%. We also give evidence that a very desirable feature for neural networks is the hyperbolicity of their Jacobian at initialization, while remaining at the edge of chaos.

## [Understanding the Eluder Dimension](#)

- Gene Li · Pritish Kamath · Dylan J Foster · Nati Srebro
- abstract@[open-review](#): We provide new insights on eluder dimension, a complexity measure that has been extensively used to bound the regret of algorithms for online bandits and reinforcement learning with function approximation. First, we study the relationship between the eluder dimension for a function class and a generalized notion of rank, defined for any monotone ``activation''  $\sigma : \mathbb{R} \rightarrow \mathbb{R}$ , which corresponds to the minimal dimension required to represent the class as a generalized linear model. It is known that when  $\sigma$  has derivatives bounded away from 0,  $\sigma$ -rank gives rise to an upper bound on eluder dimension for any function class; we show however that eluder dimension can be exponentially smaller than  $\sigma$ -rank. We also show that the condition on the derivative is necessary; namely, when  $\sigma$  is the  $\text{relu}$  activation, the eluder dimension can be exponentially larger than  $\sigma$ -rank. For Boolean-valued function classes, we obtain a characterization of the eluder dimension in terms of star number and threshold dimension, quantities which are relevant in active learning and online learning respectively.

## [Meta-Query-Net: Resolving Purity-Informativeness Dilemma in Open-set Active Learning](#)

- Dongmin Park · Yooju Shin · Jihwan Bang · Youngjun Lee · Hwanjun Song · Jae-Gil Lee
- abstract@[open-review](#): Unlabeled examples awaiting annotations contain open-set noise inevitably. A few active learning studies have attempted to deal with this open-set noise in active learning by filtering out the noisy examples. However, because focusing on the purity of examples in a query set leads to overlooking the informativeness of the examples, the best balancing of purity and informativeness remains as an important question. In this paper, to solve this purity-informativeness dilemma in open set active learning, we propose a novel Meta-Query-Net (MQ-Net) that adaptively finds the best balancing between the two factors. Specifically, by leveraging the multi-round property of active learning, we train MQ-Net using a query set without an additional validation set. Furthermore, a clear dominance relationship between unlabeled examples is effectively captured by MQ-Net through a novel skyline regularization. Extensive experiments on multiple open-set active learning scenarios demonstrate that the proposed MQ-Net achieves 20.14% improvement in terms of accuracy, compared with the state-of-the-art methods.

## [CalFAT: Calibrated Federated Adversarial Training with Label Skewness](#)

- Chen Chen · Yuchen Liu · Xingjun Ma · Lingjuan Lyu
- abstract@[open-review](#): Recent studies have shown that, like traditional machine learning, federated learning (FL) is also vulnerable to adversarial attacks. To improve the adversarial robustness of FL, few federated adversarial training (FAT) methods have been proposed to apply adversarial training locally before global aggregation. Although these methods demonstrate promising results on independent identically distributed (IID) data, they suffer from training instability issues on non-IID data with label skewness, resulting in much degraded natural accuracy. This tends to hinder the application of FAT in real-world applications where the label distribution across the clients is often skewed. In this paper, we study the problem of FAT under label skewness, and firstly reveal one root cause of the training instability and natural accuracy degradation issues: skewed labels lead to non-identical class probabilities and heterogeneous local models. We then propose a Calibrated FAT (CalFAT) approach to tackle the instability issue by calibrating the logits adaptively to balance the classes. We show both theoretically and empirically that the optimization of CalFAT leads to homogeneous local models across the clients and better convergence point.

## [Lazy and Fast Greedy MAP Inference for Determinantal Point Process](#)

- Shinichi Hemmi · Taihei Oki · Shinsaku Sakaue · Kaito Fujii · Satoru Iwata
- abstract@[open-review](#): The maximum a posteriori (MAP) inference for determinantal point processes (DPPs) is crucial for selecting diverse items in many machine learning applications. Although DPP MAP inference is NP-hard, the greedy algorithm often finds high-quality solutions, and many researchers have studied its efficient implementation. One classical and practical method is the lazy greedy algorithm, which is applicable to general submodular function maximization, while a recent fast greedy algorithm based on the Cholesky factorization is more efficient for DPP MAP inference. This paper presents how to combine the ideas of lazy and fast, which have been considered incompatible in the literature. Our lazy and fast greedy algorithm achieves almost the same time complexity as the current best one and runs faster in practice. The idea of ``lazy + fast'' is extendable to other greedy-type algorithms. We also give a fast version of the double greedy algorithm for unconstrained DPP MAP inference. Experiments validate the effectiveness of our acceleration ideas.

## [Label Noise in Adversarial Training: A Novel Perspective to Study Robust Overfitting](#)

- Chengyu Dong · Liyuan Liu · Jingbo Shang
- abstract@[open-review](#): We show that label noise exists in adversarial training. Such label noise is due to the mismatch between the true label distribution of adversarial examples and the label inherited from clean examples — the true label distribution is distorted by the adversarial perturbation, but is neglected by the common practice that inherits labels from clean examples. Recognizing label noise sheds insights on the prevalence of robust overfitting in adversarial training, and explains its intriguing dependence on perturbation radius and data quality. Also, our label noise perspective aligns well with our observations of the epoch-wise double descent in adversarial training. Guided by our analyses, we proposed a method to automatically calibrate the label to address the label noise and robust overfitting. Our method achieves consistent performance improvements across various models and datasets without introducing new hyper-parameters or additional tuning.

## [Weakly Supervised Representation Learning with Sparse Perturbations](#)

- Kartik Ahuja · Jason Hartford · Yoshua Bengio
- abstract@[open-review](#): The theory of representation learning aims to build methods that provably invert the data generating process with minimal domain knowledge or any source of supervision. Most prior approaches require strong distributional assumptions on the latent variables and weak supervision (auxiliary information such as timestamps) to provide provable identification guarantees. In this work, we show that if one has weak supervision from observations generated by sparse perturbations of the latent variables—e.g. images in a reinforcement learning environment where actions move individual sprites—identification is achievable under unknown continuous latent distributions. We show that if the perturbations are applied only on mutually exclusive blocks of latents, we identify the latents up to those blocks. We also show that if these perturbation blocks overlap, we identify latents up to the smallest blocks shared across perturbations. Consequently, if there are blocks that intersect in one latent variable only, then such latents are identified up to permutation and scaling. We propose a natural estimation procedure based on this theory and illustrate it on low-dimensional synthetic and image-based experiments.

## [A Character-Level Length Control Algorithm for Non-Autoregressive Sentence Summarization](#)

- Puyuan Liu · Xiang Zhang · Lili Mou
- abstract@[open-review](#): Sentence summarization aims at compressing a long sentence into a short one that keeps the main gist, and has extensive real-world applications such as headline generation. In previous work, researchers have developed various approaches to improve the ROUGE score, which is the main evaluation metric for summarization, whereas controlling the summary length has not drawn much attention. In our work, we address a new problem of explicit character-level length control for summarization, and propose a dynamic programming algorithm based on the Connectionist Temporal Classification (CTC) model. Results show that our approach not only achieves higher ROUGE scores but also yields more complete sentences.

## [Toward Robust Spiking Neural Network Against Adversarial Perturbation](#)

- LING LIANG · Kaidi Xu · Xing Hu · Lei Deng · Yuan Xie
- abstract@[open-review](#): As spiking neural networks (SNNs) are deployed increasingly in real-world efficiency critical applications, the security concerns in SNNs attract more attention. Currently, researchers have already demonstrated an SNN can be attacked with adversarial examples. How to build a robust SNN becomes an urgent issue. Recently, many studies apply certified training in artificial neural networks (ANNs), which can improve the robustness of an NN model promise. However, existing certifications cannot transfer to SNNs directly because of the distinct neuron behavior and input formats for SNNs. In this work, we first design S-IBP and S-CROWN that tackle the non-linear functions in SNNs' neuron modeling. Then, we formalize the boundaries for both digital and spike inputs. Finally, we demonstrate the efficiency of our proposed robust training method in different datasets and model architectures. Based on our experiment, we can achieve a maximum \$37.7\%\$ attack error reduction with \$3.7\%\$ original accuracy loss. To the best of our knowledge, this is the first analysis on robust training of SNNs.

## [Risk-Driven Design of Safety-Critical Perception Systems](#)

- Anthony Corso · Sydney Katz · Craig Innes · Xin Du · Subramanian Ramamoorthy · Mykel J Kochenderfer
- abstract@[open-review](#): Modern autonomous systems rely on perception modules to process complex sensor measurements into state estimates. These estimates are then passed to a controller, which uses them to make safety-critical decisions. It is therefore important that we design perception systems to minimize errors that reduce the overall safety of the system. We develop a risk-driven approach to designing perception systems that accounts for the effect of perceptual errors on the performance of the fully-integrated, closed-loop system. We formulate a risk function to quantify the effect of a given perceptual error on overall safety, and show how we can use it to design safer perception systems by including a risk-dependent term in the loss function and generating training data in risk-sensitive regions. We evaluate our techniques on a realistic vision-based aircraft detect and avoid application and show that risk-driven design reduces collision risk by 37% over a baseline system.

## [Distributed Inverse Constrained Reinforcement Learning for Multi-agent Systems](#)

- Shicheng Liu · Minghui Zhu
- abstract@[open-review](#): This paper considers the problem of recovering the policies of multiple interacting experts by estimating their reward functions and constraints where the demonstration data of the experts is distributed to a group of learners. We formulate this problem as a distributed bi-level optimization problem and propose a novel bi-level ``distributed inverse constrained reinforcement learning'' (D-ICRL) algorithm that allows the learners to collaboratively estimate the constraints in the outer loop and learn the corresponding policies and reward functions in the inner loop from the distributed demonstrations through intermittent communications. We formally guarantee that the distributed learners asymptotically achieve consensus which belongs to the set of stationary points of the bi-level optimization problem.

## [Flatten the Curve: Efficiently Training Low-Curvature Neural Networks](#)

- Suraj Srinivas · Kyle Matoba · Himabindu Lakkaraju · François Fleuret
- abstract@[open-review](#): Deep neural networks suffer from adversarial vulnerability and poor interpretability owing to their high degree of non-linearity, which manifests as a large model curvature. Curvature encodes non-linearity typically via Hessian norms. Low-curvature neural network models can help avoid these problems, but existing methods are expensive to train and often sacrifice predictive accuracy. In this work, we demonstrate low-curvature neural networks (LCNNs) that obtain lower curvature than standard models while exhibiting similar predictive performance, and only marginally increased training time. To achieve this, we minimize a data-independent upper bound on the curvature of a neural network, which decomposes overall curvature in terms of curvatures and slopes of its constituent layers. To efficiently minimize this bound, we introduce two novel architectural components: first, a non-linearity called centered-softplus that is a stable variant of the softplus non-linearity, and second, a Lipschitz-constrained batch normalization layer. Our experiments show that LCNNs have lower curvature, more stable gradients and increased off-the-shelf adversarial robustness when compared to their standard high-curvature counterparts, all without affecting predictive performance. Our approach is easy to use and can be readily incorporated into existing neural network models to remove their excess curvature.

## [Self-explaining deep models with logic rule reasoning](#)

- Seungeon Lee · Xiting Wang · Sungwon Han · Eunji Lee · Xiaoyuan Yi · Xing Xie · Meeyoung Cha
- abstract@[open-review](#): We present a framework for integrating self-explaining capabilities into a given deep model to achieve high prediction performance and human precision. Human precision means that most model decisions are coherent with human decision logic, allowing users to quickly understand and identify a small fraction of problematic model behavior. We propose two desirable properties for ensuring high human precision and demonstrate that logic rule explanations naturally satisfy them, while also possessing the expressive power required for good predictive performance. Then, using logic rules, we propose a framework for a deep model to predict and explain. Our method does not require predefined logic rule sets, and can be learned in a differentiable way with widely-used deep learning modules. Extensive experiments show that our method achieves high prediction performance and human precision, as well as being naturally robust to noisy labels.

## [A Mean-Field Game Approach to Cloud Resource Management with Function Approximation](#)

- Weichao Mao · Haoran Qiu · Chen Wang · Hubertus Franke · Zbigniew Kalbarczyk · Ravishankar Iyer · Tamer Basar

- abstract@[open-review](#): Reinforcement learning (RL) has gained increasing popularity for resource management in cloud services such as serverless computing. As self-interested users compete for shared resources in a cluster, the multi-tenancy nature of serverless platforms necessitates multi-agent reinforcement learning (MARL) solutions, which often suffer from severe scalability issues. In this paper, we propose a mean-field game (MFG) approach to cloud resource management that is scalable to a large number of users and applications and incorporates function approximation to deal with the large state-action spaces in real-world serverless platforms. Specifically, we present an online natural actor-critic algorithm for learning in MFGs compatible with various forms of function approximation. We theoretically establish its finite-time convergence to the regularized Nash equilibrium under linear function approximation and softmax parameterization. We further implement our algorithm using both linear and neural-network function approximations, and evaluate our solution on an open-source serverless platform, OpenWhisk, with real-world workloads from production traces. Experimental results demonstrate that our approach is scalable to a large number of users and significantly outperforms various baselines in terms of function latency and resource utilization efficiency.

## [Causal Disentanglement for Time Series](#)

- Weiran Yao · Guangyi Chen · Kun Zhang
- abstract@[open-review](#): Recently in the field of nonlinear Independent Component Analysis (ICA), strong identifiability results for disentanglement have been established by using certain side information, such as class labels, or history information for time series, in addition to independence. However, most existing work is constrained by functional form assumptions such as stationary independent sources or further with linear transitions, and distribution assumptions such as exponential family distribution. It is unknown whether the underlying latent processes and their causal relations are identifiable if they have arbitrary, nonparametric causal influences in between. We propose a principled framework called LCD-NM to recover time-delayed latent causal variables and identify their relations from measured temporal data under stationary environments and under different distribution shifts. Specifically, the framework factorizes unknown distribution shifts into transition distribution changes caused by fixed dynamics and time-varying latent causal relations, and by global changes in observation. We establish the identifiability theories of nonparametric latent causal processes from their nonlinear mixtures under fixed dynamics and analyze how distribution changes can further benefit the identifiability. Through experiments, we show that time-delayed latent causal influences are reliably identified and that our approach considerably outperforms existing baselines that do not properly exploit this modular representation of changes. Our results demonstrate that disentanglement in time-series settings seems promising both in stationary environments and general nonstationary environments, in which the latent processes have nonparametric causal influences in between.

## [FlowHMM: Flow-based continuous hidden Markov models](#)

- Paweł Lorek · Rafał Nowak · Tomasz Trzcinski · Maciej Zieba
- abstract@[open-review](#): Continuous hidden Markov models (HMMs) assume that observations are generated from a mixture of Gaussian densities, limiting their ability to model more complex distributions. In this work, we address this shortcoming and propose novel continuous HMM models, dubbed FlowHMMs, that allow to learn general continuous observation densities without constraining them to follow a Gaussian distribution or their mixtures. To that end, we leverage deep flow-based architectures that model complex, non-Gaussian functions and propose two variants of training FlowHMM model. The first one, based on an expectation-maximization (EM) technique, can be applied directly to continuous multidimensional data, yet its application to larger data sequences remains computationally expensive. Therefore, we also present a second approach to training our FlowHMM that relies on the co-occurrence matrix of discretized observations and considers the joint distribution of pairs of co-observed values, hence rendering the training time independent of the training sequence length. As a result, we obtain a model that can be flexibly adapted to the characteristics and dimensionality of the data. We perform a variety of experiments in which we compare both training strategies with a baseline of Gaussian mixture models. We show that in terms of quality of the recovered probability distribution, accuracy of prediction of hidden states, and likelihood of unseen data, our approach outperforms the standard Gaussian methods.

## [Anonymized Histograms in Intermediate Privacy Models](#)

- Badh Ghazi · Pritish Kamath · Ravi Kumar · Pasin Manurangsi
- abstract@[open-review](#): We study the problem of privately computing the  $\text{anonymized histogram}$  (a.k.a.  $\text{unattributed histogram}$ ), which is defined as the histogram without item labels. Previous works have provided algorithms with  $\ell_1$  and  $\ell_2^2$ -errors of  $O(\varepsilon \sqrt{n})$  in the central model of differential privacy (DP). In this work, we provide an algorithm with a nearly matching error guarantee of  $\tilde{O}(\varepsilon \sqrt{n})$  in the shuffle DP and pan-private models. Our algorithm is very simple: it just post-processes the discrete Laplace-noised histogram! Using this algorithm as a subroutine, we show applications in privately estimating symmetric properties of distributions such as entropy.

## [Maximizing and Satisficing in Multi-armed Bandits with Graph Information](#)

- Parth Thaker · Mohit Malu · Nikhil Rao · Gautam Dasarathy
- abstract@[open-review](#): Pure exploration in multi-armed bandits has emerged as an important framework for modeling decision making and search under uncertainty. In modern applications however, one is often faced with a tremendously large number of options and even obtaining one observation per option may be too costly rendering traditional pure exploration algorithms ineffective. Fortunately, one often has access to similarity relationships amongst the options that can be leveraged. In this paper, we consider the pure exploration problem in stochastic multi-armed bandits where the similarities between the arms is captured by a graph and the rewards may be represented as a smooth signal on this graph. In particular, we consider the problem of finding the arm with the maximum reward (i.e., the maximizing problem) or one that has sufficiently high reward (i.e., the satisficing problem) under this model. We propose novel algorithms GRUB (GRaph based UcB) and zeta-GRUB for these problems and provide theoretical characterization of their performance which specifically elicits the benefit of the graph side information. We also prove a lower bound on the data requirement that shows a large class of problems where these algorithms are near-optimal. We complement our theory with experimental results that show the benefit of capitalizing on such side information.

## [Understanding and Improving Robustness of Vision Transformers through Patch-based Negative Augmentation](#)

- Yao Qin · Chiyuan Zhang · Ting Chen · Balaji Lakshminarayanan · Alex Beutel · Xuezhi Wang
- abstract@[open-review](#): We investigate the robustness of vision transformers (ViTs) through the lens of their special patch-based architectural structure, i.e., they process an image as a sequence of image patches. We find that ViTs are surprisingly insensitive to patch-based transformations, even when the transformation largely destroys the original semantics and makes the image unrecognizable by humans. This indicates that ViTs heavily use features that survived such transformations but are generally not indicative of the semantic class to humans. Further investigations show that these features are useful but non-robust, as ViTs trained on them can achieve high in-distribution accuracy, but break down under distribution shifts. From this understanding, we ask: can training the model to rely less on these features improve ViT robustness and out-of-distribution performance? We use the images transformed with our patch-based operations as negatively augmented views and offer losses to regularize the training away from using non-robust features. This is a complementary view to existing research that mostly focuses on augmenting inputs with semantic-preserving transformations to enforce models' invariance. We show that patch-based negative augmentation consistently improves robustness of ViTs on ImageNet based robustness benchmarks across 20+ different experimental settings. Furthermore, we find our patch-based negative augmentation are complementary to traditional (positive) data augmentation techniques and batch-based negative examples in contrastive learning. All the code will be open-sourced.

## [Markovian Interference in Experiments](#)

- Vivek Farias · Andrew Li · Tianyi Peng · Andrew Zheng
- abstract@[open-review](#): We consider experiments in dynamical systems where interventions on some experimental units impact other units through a limiting constraint (such as a limited supply of products). Despite outsize practical importance, the best estimators for this ‘Markovian’ interference problem are largely heuristic in nature, and their bias is not well understood. We formalize the problem of inference in such experiments as one of policy evaluation. Off-policy estimators, while unbiased, apparently incur a large penalty in variance relative to state-of-the-art heuristics. We introduce an on-policy estimator: the Differences-In-Q’s (DQ) estimator. We show that the DQ estimator can in general have exponentially smaller variance than off-policy evaluation. At the same time, its bias is second order in the impact of the intervention. This yields a striking bias-variance tradeoff so that the DQ estimator effectively dominates state-of-the-art alternatives. Our empirical evaluation includes a set of experiments on a city-scale ride-hailing simulator.

## Lifting Weak Supervision To Structured Prediction

- Harit Vishwakarma · Frederic Sala
- abstract@[open-review](#): Weak supervision (WS) is a rich set of techniques that produce pseudolabels by aggregating easily obtained but potentially noisy label estimates from various sources. WS is theoretically well-understood for binary classification, where simple approaches enable consistent estimation of pseudolabel noise rates. Using this result, it has been shown that downstream models trained on the pseudolabels have generalization guarantees nearly identical to those trained on clean labels. While this is exciting, users often wish to use WS for \emph{structured prediction}, where the output space consists of more than a binary or multi-class label set: e.g. rankings, graphs, manifolds, and more. Do the favorable theoretical properties of WS for binary classification lift to this setting? We answer this question in the affirmative for a wide range of scenarios. For labels taking values in a finite metric space, we introduce techniques new to weak supervision based on pseudo-Euclidean embeddings and tensor decompositions, providing a nearly-consistent noise rate estimator. For labels in constant-curvature Riemannian manifolds, we introduce new invariants that also yield consistent noise rate estimation. In both cases, when using the resulting pseudolabels in concert with a flexible downstream model, we obtain generalization guarantees nearly identical to those for models trained on clean data. Several of our results, which can be viewed as robustness guarantees in structured prediction with noisy labels, may be of independent interest.

## Masked Autoencoders that Listen

- Po-Yao Huang · Hu Xu · Juncheng Li · Alexei Baevski · Michael Auli · Wojciech Galuba · Florian Metze · Christoph Feichtenhofer
- abstract@[open-review](#): This paper studies a simple extension of image-based Masked Autoencoders (MAE) to self-supervised representation learning from audio spectrograms. Following the Transformer encoder-decoder design in MAE, our Audio-MAE first encodes audio spectrogram patches with a high masking ratio, feeding only the non-masked tokens through encoder layers. The decoder then re-orders and decodes the encoded context padded with mask tokens, in order to reconstruct the input spectrogram. We find it beneficial to incorporate local window attention in the decoder, as audio spectrograms are highly correlated in local time and frequency bands. We then fine-tune the encoder with a lower masking ratio on target datasets. Empirically, Audio-MAE sets new state-of-the-art performance on six audio and speech classification tasks, outperforming other recent models that use external supervised pre-training. Our code and models will be available.

## Unsupervised Point Cloud Completion and Segmentation by Generative Adversarial Autoencoding Network

- Changfeng Ma · Yang Yang · Jie Guo · Fei Pan · Chongjun Wang · Yanwen Guo
- abstract@[open-review](#): Most existing point cloud completion methods assume the input partial point cloud is clean, which is not practical in practice, and are generally based on supervised learning. In this paper, we present an unsupervised generative adversarial autoencoding network, named UGAAN, which completes the partial point cloud contaminated by surroundings from real scenes and cutouts the object simultaneously, only using artificial CAD models as assistance. The generator of UGAAN learns to predict the complete point clouds on real data from both the discriminator and the autoencoding process of artificial data. The latent codes from generator are also fed to discriminator which makes encoder only extract object features rather than noises. We also devise a refiner for generating better complete cloud with a segmentation module to separate the object from background. We train our UGAAN with one real scene dataset and evaluate it with the other two. Extensive experiments and visualization demonstrate our superiority, generalization and robustness. Comparisons against the previous method show that our method achieves the state-of-the-art performance on unsupervised point cloud completion and segmentation on real data.

## Make Some Noise: Reliable and Efficient Single-Step Adversarial Training

- Pau de Jorge Aranda · Adel Bibi · Riccardo Volpi · Amartya Sanyal · Philip Torr · Gregory Rogez · Puneet Dokania
- abstract@[open-review](#): Recently, Wong et al. (2020) showed that adversarial training with single-step FGSM leads to a characteristic failure mode named catastrophic overfitting (CO), in which a model becomes suddenly vulnerable to multi-step attacks. Experimentally they showed that simply adding a random perturbation prior to FGSM (RS-FGSM) could prevent CO. However, Andriushchenko & Flammarion (2020) observed that RS-FGSM still leads to CO for larger perturbations, and proposed a computationally expensive regularizer (GradAlign) to avoid it. In this work, we methodically revisit the role of noise and clipping in single-step adversarial training. Contrary to previous intuitions, we find that using a stronger noise around the clean sample combined with \textit{not} clipping is highly effective in avoiding CO for large perturbation radii. We then propose Noise-FGSM (N-FGSM) that, while providing the benefits of single-step adversarial training, does not suffer from CO. Empirical analyses on a large suite of experiments show that N-FGSM is able to match or surpass the performance of previous state-of-the-art GradAlign while achieving 3\\$times\\$ speed-up.

## BEVFusion: A Simple and Robust LiDAR-Camera Fusion Framework

- Tingting Liang · Hongwei Xie · Kaicheng Yu · Zhongyu Xia · Zhiwei Lin · Yongtao Wang · Tao Tang · Bing Wang · Zhi Tang
- abstract@[open-review](#): Fusing the camera and LiDAR information has become a de-facto standard for 3D object detection tasks. Current methods rely on point clouds from the LiDAR sensor as queries to leverage the feature from the image space. However, people discovered that this underlying assumption makes the current fusion framework infeasible to produce any prediction when there is a LiDAR malfunction, regardless of minor or major. This fundamentally limits the deployment capability to realistic autonomous driving scenarios. In contrast, we propose a surprisingly simple yet novel fusion framework, dubbed BEVFusion, whose camera stream does not depend on the input of LiDAR data, thus addressing the downside of previous methods. We empirically show that our framework surpasses the state-of-the-art methods under the normal training settings. Under the robustness training settings that simulate various LiDAR malfunctions, our framework significantly surpasses the state-of-the-art methods by 15.7% to 28.9% mAP. To the best of our knowledge, we are the first to handle realistic LiDAR malfunction and can be deployed to realistic scenarios without any post-processing procedure.

## Regret Bounds for Multilabel Classification in Sparse Label Regimes

- RÃ³bert Busa-Fekete · Heejin Choi · Krzysztof Dembczynski · Claudio Gentile · Henry Reeve · Balazs Szorenyi
- abstract@[open-review](#): Multi-label classification (MLC) has wide practical importance, but the theoretical understanding of its statistical properties is still limited. As an attempt to fill this gap, we thoroughly study upper and lower regret bounds for two canonical MLC performance measures, Hamming loss and Precision@\$\\kappa\$. We consider two different statistical and algorithmic settings, a non-parametric setting tackled by plug-in classifiers ‘a la \$k\$-nearest neighbors, and a parametric one tackled by empirical risk minimization operating on surrogate loss functions. For both, we analyze the interplay between a natural MLC variant of the low noise assumption, widely studied in binary classification, and the label sparsity, the latter being a natural

property of large-scale MLC problems. We show that those conditions are crucial in improving the bounds, but the way they are tangled is not obvious, and also different across the two settings.

## [Multi-agent Covering Option Discovery based on Kronecker Product of Factor Graphs](#)

- Jiayu Chen · Jingdi Chen · Tian Lan · Vaneet Aggarwal
- abstract@[open-review](#): Covering option discovery has been developed to improve the exploration of RL in single-agent scenarios with sparse reward signals, through connecting the most distant states in the embedding space provided by the Fiedler vector of the state transition graph. Given that joint state space grows exponentially with the number of agents in multi-agent systems, existing researches still relying on single-agent option discovery either become prohibitive or fail to directly discover joint options that improve the connectivity of the joint state space. In this paper, we show how to directly compute multi-agent options with collaborative exploratory behaviors while still enjoying the ease of decomposition. Our key idea is to approximate the joint state space as the Kronecker product of individual agents' state spaces, based on which we can directly estimate the Fiedler vector of the joint state space using the Laplacian spectrum of individual agents' transition graphs. We also extend our method to tasks with infinite-scale state space by estimating eigenfunctions through deep representation learning techniques. The evaluation on multi-agent tasks built with simulators like Mujoco, shows that the proposed algorithm can successfully identify multi-agent options, and significantly outperforms the state-of-the-art. Codes are available at: <https://anonymous.4open.science/r/NIPS2022exp>.

## [Neural Transmitted Radiance Fields](#)

- Chengxuan Zhu · Renjie Wan · Boxin Shi
- abstract@[open-review](#): Neural radiance fields (NeRF) have brought tremendous progress to novel view synthesis. Though NeRF enables the rendering of subtle details in a scene by learning from a dense set of images, it also reconstructs the undesired reflections when we capture images through glass. As a commonly observed interference, the reflection would undermine the visibility of the desired transmitted scene behind glass by occluding the transmitted light rays. In this paper, we aim at addressing the problem of rendering novel transmitted views given a set of reflection-corrupted images. By introducing the transmission encoder and recurring edge constraints as guidance, our neural transmitted radiance fields can resist such reflection interference during rendering and reconstruct high-fidelity results even under sparse views. The proposed method achieves superior performance from the experiments on a newly collected dataset compared with state-of-the-art methods.

## [Deep Differentiable Logic Gate Networks](#)

- Felix Petersen · Christian Borgelt · Hilde Kuehne · Oliver Deussen
- abstract@[open-review](#): Recently, research has increasingly focused on developing efficient neural network architectures. In this work, we explore logic gate networks for machine learning tasks by learning combinations of logic gates. These networks comprise logic gates such as AND" and XOR", which allows very fast execution. The difficulty in learning logic gate networks is that they are conventionally non-differentiable and therefore do not allow training with gradient descent. We propose differentiable logic gate networks, an architecture that combines real-valued logics and a continuously parameterized relaxation of the network. Differentiable logic gate networks allow optimizing the composition of logic gates via gradient descent.

## [Neural Basis Models for Interpretability](#)

- Filip Radenovic · Abhimanyu Dubey · Dhruv Mahajan
- abstract@[open-review](#): Due to the widespread use of complex machine learning models in real-world applications, it is becoming critical to explain model predictions. However, these models are typically black-box deep neural networks, explained post-hoc via methods with known faithfulness limitations. Generalized Additive Models (GAMs) are an inherently interpretable class of models that address this limitation by learning a non-linear shape function for each feature separately, followed by a linear model on top. However, these models are typically difficult to train, require numerous parameters, and are difficult to scale. We propose an entirely new subfamily of GAMs that utilizes basis decomposition of shape functions. A small number of basis functions are shared among all features, and are learned jointly for a given task, thus making our model scale much better to large-scale data with high-dimensional features, especially when features are sparse. We propose an architecture denoted as the Neural Basis Model (NBM) which uses a single neural network to learn these bases. On a variety of tabular and image datasets, we demonstrate that for interpretable machine learning, NBMs are the new state-of-the-art in accuracy, model size, and, throughput. Source code will be made available upon acceptance.

## [On Divergence Measures for Bayesian Pseudocoresets](#)

- Balhae Kim · Jungwon Choi · Seanie Lee · Yoonho Lee · Jung-Woo Ha · Juho Lee
- abstract@[open-review](#): A Bayesian pseudocoreset is a small synthetic dataset for which the posterior over parameters approximates that of the original dataset. While promising, the scalability of Bayesian pseudocoresets is not yet validated in large-scale problems such as image classification with deep neural networks. On the other hand, dataset distillation methods similarly construct a small dataset such that the optimization with the synthetic dataset converges to a solution similar to optimization with full data. Although dataset distillation has been empirically verified in large-scale settings, the framework is restricted to point estimates, and their adaptation to Bayesian inference has not been explored. This paper casts two representative dataset distillation algorithms as approximations to methods for constructing pseudocoresets by minimizing specific divergence measures: reverse KL divergence and Wasserstein distance. Furthermore, we provide a unifying view of such divergence measures in Bayesian pseudocoreset construction. Finally, we propose a novel Bayesian pseudocoreset algorithm based on minimizing forward KL divergence. Our empirical results demonstrate that the pseudocoresets constructed from these methods reflect the true posterior even in large-scale Bayesian inference problems.

## [Test-Time Prompt Tuning for Zero-Shot Generalization in Vision-Language Models](#)

- Manli Shu · Chaowei Xiao · Weili Nie · De-An Huang · Zhiding Yu · Tom Goldstein · Anima Anandkumar
- abstract@[open-review](#): Pre-trained vision-language models (e.g., CLIP) have shown promising zero-shot generalization in many downstream tasks with properly designed text prompts. Instead of relying on hand-engineered prompts, recent works learn prompts using the training data from downstream tasks. While effective, training on domain-specific data reduces a model's generalization capability to unseen new domains. In this work, we propose test-time prompt tuning (TPT), a method that can learn adaptive prompts on the fly with a single test sample. TPT optimizes the prompt by minimizing the entropy with confidence selection so that the model has consistent predictions across different augmented views of each test sample. In evaluating generalization to natural distribution shifts, TPT improves the zero-shot top-1 accuracy of CLIP by 3.6% on average, surpassing previous prompt tuning approaches that require additional task-specific training data. In evaluating cross-dataset generalization with unseen categories, TPT performs on par with the state-of-the-art approaches that use additional training data.

## [Exact Solutions of a Deep Linear Network](#)

- Liu Ziyin · Botao Li · Xiangming Meng
- abstract@[open-review](#): This work finds the analytical expression of the global minima of a deep linear network with weight decay and stochastic neurons, a fundamental model for understanding the landscape of neural networks. Our result implies that zero is a special point in deep neural network architecture. We show that weight decay strongly interacts with the model architecture and can create bad minima at zero in a network with more than \$1\$

hidden layer, qualitatively different from a network with only \$1\$ hidden layer. Practically, our result implies that common deep learning initialization methods are insufficient to ease the optimization of neural networks in general.

## [Learning to Scaffold: Optimizing Model Explanations for Teaching](#)

- Patrick Fernandes · Marcos Treviso · Danish Pruthi · André Martins · Graham Neubig
- abstract@[open-review](#): Modern machine learning models are opaque, and as a result there is a burgeoning academic subfield on methods that explain these models' behavior. However, what is the precise goal of providing such explanations, and how can we demonstrate that explanations achieve this goal? Some research argues that explanations should help teach a student (either human or machine) to simulate the model being explained, and that the quality of explanations can be measured by the simulation accuracy of students on unexplained examples. In this work, leveraging meta-learning techniques, we extend this idea to improve the quality of the explanations themselves, specifically by optimizing explanations such that student models more effectively learn to simulate the original model. We train models on three natural language processing and computer vision tasks, and find that students trained with explanations extracted with our framework are able to simulate the teacher significantly more effectively than ones produced with previous methods. Through human annotations and a user study, we further find that these learned explanations more closely align with how humans would explain the required decisions in these tasks. Our code is available at <https://anonymous.4open.science/r/learning-scaffold-5BEB>

## [Maximum Likelihood Training of Implicit Nonlinear Diffusion Model](#)

- Dongjun Kim · Byeonghu Na · Se Jung Kwon · Dongsoo Lee · Wanmo Kang · Il-chul Moon
- abstract@[open-review](#): Whereas diverse variations of diffusion models exist, expanding the linear diffusion into a nonlinear diffusion process is investigated only by a few works. The nonlinearity effect has been hardly understood, but intuitively, there would be more promising diffusion patterns to optimally train the generative distribution towards the data distribution. This paper introduces such a data-adaptive and nonlinear diffusion process for score-based diffusion models. The proposed Implicit Nonlinear Diffusion Model (INDM) learns the nonlinear diffusion process by combining a normalizing flow and a diffusion process. Specifically, INDM implicitly constructs a nonlinear diffusion on the \textit{data space} by leveraging a linear diffusion on the \textit{latent space} through a flow network. This flow network is the key to forming a nonlinear diffusion as the nonlinearity fully depends on the flow network. This flexible nonlinearity is what improves the learning curve of INDM to nearly MLE training, compared against the non-MLE training of DDPM++, which turns out to be a special case of INDM with the identity flow. Also, training the nonlinear diffusion empirically yields a sampling-friendly latent diffusion that the sample trajectory of INDM is closer to an optimal transport than the trajectories of previous research. In experiments, INDM achieves the state-of-the-art FID on CelebA.

## [Relation-Constrained Decoding for Text Generation](#)

- Xiang Chen · Zhixian Yang · Xiaojun Wan
- abstract@[open-review](#): The dominant paradigm for neural text generation nowadays is seq2seq learning with large-scale pretrained language models. However, it is usually difficult to manually constrain the generation process of these models. Prior studies have introduced Lexically Constrained Decoding (LCD) to ensure the presence of pre-specified words or phrases in the output. However, simply applying lexical constraints has no guarantee of the grammatical or semantic relations between words. Thus, more elaborate constraints are needed. To this end, we first propose a new constrained decoding scenario named Relation-Constrained Decoding (RCD), which requires the model's output to contain several given word pairs with respect to the given relations between them. For this scenario, we present a novel plug-and-play decoding algorithm named RElation-guided probability Surgery and bEam ALlocation (RESEAL), which can handle different categories of relations, e.g., syntactical relations or factual relations. Moreover, RESEAL can adaptively "reseal" the relations to form a high-quality sentence, which can be applied to the inference stage of any autoregressive text generation model. To evaluate our method, we first construct an RCD benchmark based on dependency relations from treebanks with annotated dependencies. Experimental results demonstrate that our approach can achieve better preservation of the input dependency relations compared to previous methods. To further illustrate the effectiveness of RESEAL, we apply our method to three downstream tasks: sentence summarization, data-to-text generation, and fact-based text editing. We observe an improvement in generation quality.

## [Guaranteed Conservation of Momentum for Learning Particle-based Fluid Dynamics](#)

- Lukas Prantl · Benjamin Ummenhofer · Vladlen Koltun · Nils Thuerey
- abstract@[open-review](#): We present a novel method for guaranteeing linear momentum in learned physics simulations. Unlike existing methods, we enforce conservation of momentum with a hard constraint, which we realize via antisymmetrical continuous convolutional layers. We combine these strict constraints with a hierarchical network architecture, a carefully constructed resampling scheme, and a training approach for temporal coherence. In combination, the proposed method allows us to substantially increase the physical accuracy of the learned simulator. In addition, the induced physics bias leads to significantly better generalization performance and makes our method more reliable in unseen test cases. We evaluate our method on a range of different, challenging fluid scenarios, and show that the proposed algorithm can learn complex dynamics while outperforming existing approaches in terms of generalization and training performance.

## [Poisson Flow Generative Models](#)

- Yilun Xu · Ziming Liu · Max Tegmark · Tommi Jaakkola
- abstract@[open-review](#): We propose a new "Poisson flow" generative model~(PFGM) that maps a uniform distribution on a high-dimensional hemisphere into any data distribution. We interpret the data points as electrical charges on the  $z=0$  hyperplane in a space augmented with an additional dimension  $z$ , generating a high-dimensional electric field (the gradient of the solution to Poisson equation). We prove that if these charges flow upward along electric field lines, their initial distribution in the  $z=0$  plane transforms into a distribution on the hemisphere of radius  $r$  that becomes uniform in the  $\rightarrow\infty$  limit. To learn the bijective transformation, we estimate the normalized field in the augmented space. For sampling, we devise a backward ODE that is anchored by the physically meaningful additional dimension: the samples hit the (unaugmented) data manifold when the  $z$  reaches zero. Experimentally, PFGM achieves current state-of-the-art performance among the normalizing flow models on CIFAR-10, with an Inception score of  $9.68$  and a FID score of  $2.48$ . It also performs on par with the state-of-the-art SDE approaches while offering  $10\times$  to  $20\times$  acceleration on image generation tasks. Additionally, PFGM appears more tolerant of estimation errors on a weaker network architecture and robust to the step size in the Euler method. The code is available at [https://github.com/Newbeeer/poisson\\_flow](https://github.com/Newbeeer/poisson_flow).

## [On Analyzing Generative and Denoising Capabilities of Diffusion-based Deep Generative Models](#)

- Kamil Deja · Anna Kuzina · Tomasz Trzcinski · Jakub Tomczak
- abstract@[open-review](#): Diffusion-based Deep Generative Models (DDGMs) offer state-of-the-art performance in generative modeling. Their main strength comes from their unique setup in which a model (the backward diffusion process) is trained to reverse the forward diffusion process, which gradually adds noise to the input signal. Although DDGMs are well studied, it is still unclear how the small amount of noise is transformed during the backward diffusion process. Here, we focus on analyzing this problem to gain more insight into the behavior of DDGMs and their denoising and generative capabilities. We observe a fluid transition point that changes the functionality of the backward diffusion process from generating a (corrupted) image from noise to denoising the corrupted image to the final sample. Based on this observation, we postulate to divide a DDGM into two parts: a denoiser and a generator. The denoiser could be parameterized by a denoising auto-encoder, while the generator is a diffusion-based model with its own set of parameters. We experimentally validate our proposition, showing its pros and cons.

## Efficiency Ordering of Stochastic Gradient Descent

- Jie Hu · Vishwaraj Doshi · Do-Young Eun
- abstract@[open-review](#): We consider the stochastic gradient descent (SGD) algorithm driven by a general stochastic sequence, including i.i.d noise and random walk on an arbitrary graph, among others; and analyze it in the asymptotic sense. Specifically, we employ the notion of ‘efficiency ordering’, a well-analyzed tool for comparing the performance of Markov Chain Monte Carlo (MCMC) samplers, for SGD algorithms in the form of Loewner ordering of covariance matrices associated with the scaled iterate errors in the long term. Using this ordering, we show that input sequences that are more efficient for MCMC sampling also lead to smaller covariance of the errors for SGD algorithms in the limit. This also suggests that an arbitrarily weighted MSE of SGD iterates in the limit becomes smaller when driven by more efficient chains. Our finding is of particular interest in applications such as decentralized optimization and swarm learning, where SGD is implemented in a random walk fashion on the underlying communication graph for cost issues and/or data privacy. We demonstrate how certain non-Markovian processes, for which typical mixing-time based non-asymptotic bounds are intractable, can outperform their Markovian counterparts in the sense of efficiency ordering for SGD. We show the utility of our method by applying it to gradient descent with shuffling and mini-batch gradient descent, reaffirming key results from existing literature under a unified framework. Empirically, we also observe efficiency ordering for variants of SGD such as accelerated SGD and Adam, open up the possibility of extending our notion of efficiency ordering to a broader family of stochastic optimization algorithms.

## Decision-Focused Learning without Decision-Making: Learning Locally Optimized Decision Losses

- Sanket Shah · Kai Wang · Bryan Wilder · Andrew Perrault · Milind Tambe
- abstract@[open-review](#): Decision-Focused Learning (DFL) is a paradigm for tailoring a predictive model to a downstream optimization task that uses its predictions in order to perform better \textit{on that specific task}. The main technical challenge associated with DFL is that it requires being able to differentiate through the optimization problem, which is difficult due to discontinuous solutions and other challenges. Past work has largely gotten around this issue by \textit{handcrafting} task-specific surrogates to the original optimization problem that provide informative gradients when differentiated through. However, the need to handcraft surrogates for each new task limits the usability of DFL. In addition, there are often no guarantees about the convexity of the resulting surrogates and, as a result, training a predictive model using them can lead to inferior local optima. In this paper, we do away with surrogates altogether and instead \textit{learn} loss functions that capture task-specific information. To the best of our knowledge, ours is the first approach that entirely replaces the optimization component of decision-focused learning with a loss that is automatically learned. Our approach (a) only requires access to a black-box oracle that can solve the optimization problem and is thus \textit{generalizable}, and (b) can be \textit{convex by construction} and so can be easily optimized over. We evaluate our approach on three resource allocation problems from the literature and find that our approach outperforms learning without taking into account task-structure in all three domains, and even hand-crafted surrogates from the literature.

## Losses Can Be Blessings: Routing Self-Supervised Speech Representations Towards Efficient Multilingual and Multitask Speech Processing

- Yonggan Fu · Yang Zhang · Kaizhi Qian · Zhifan Ye · Zhongzhi Yu · Cheng-I Jeff Lai · Yingyan Lin
- abstract@[open-review](#): Self-supervised learning (SSL) for rich speech representations has achieved empirical success in low-resource Automatic Speech Recognition (ASR) and other speech processing tasks, which can mitigate the necessity of a large amount of transcribed speech and thus has driven a growing demand for on-device ASR and other speech processing. However, advanced speech SSL models have become increasingly large, which contradicts the limited on-device resources. This gap could be more severe in multilingual/multitask scenarios requiring simultaneously recognizing multiple languages or executing multiple speech processing tasks. Additionally, strongly overparameterized speech SSL models tend to suffer from overfitting when being finetuned on low-resource speech corpus. This work aims to enhance the practical usage of speech SSL models towards a win-win in both enhanced efficiency and alleviated overfitting via our proposed S\$^3\$-Router framework, which for the first time discovers that simply discarding no more than 10% of model weights via only finetuning model connections of speech SSL models can achieve better accuracy over standard weight finetuning on downstream speech processing tasks. More importantly, S\$^3\$-Router can serve as an all-in-one technique to enable (1) a new finetuning scheme, (2) an efficient multilingual/multitask solution, (3) a state-of-the-art pruning technique, and (4) a new tool to quantitatively analyze the learned speech representation. All codes will be released upon acceptance.

## Mirror Descent with Relative Smoothness in Measure Spaces, with application to Sinkhorn and EM

- Pierre-Cyril Aubin-Frankowski · Anna Korba · Flavien LÃ©ger
- abstract@[open-review](#): Many problems in machine learning can be formulated as optimizing a convex functional over a space of measures. This paper studies the convergence of the mirror descent algorithm in this infinite-dimensional setting. Defining Bregman divergences through directional derivatives, we derive the convergence of the scheme for relatively smooth and strongly convex pairs of functionals. Applying our result to joint distributions and the Kullback-Leibler (KL) divergence, we show that Sinkhorn's primal iterations for entropic optimal transport in the continuous setting correspond to a mirror descent, and we obtain a new proof of its (sub)linear convergence. We also show that Expectation Maximization (EM) can always formally be written as a mirror descent, and, when optimizing on the latent distribution while fixing the mixtures, we derive sublinear rates of convergence.

## Interaction-Grounded Learning with Action-inclusive Feedback

- Tengyang Xie · Akanksha Saran · Dylan J Foster · Lekan Molu · Ida Momennejad · Nan Jiang · Paul Mineiro · John Langford
- abstract@[open-review](#): Consider the problem setting of Interaction-Grounded Learning (IGL), in which a learner’s goal is to optimally interact with the environment with no explicit reward to ground its policies. The agent observes a context vector, takes an action, and receives a feedback vector, using this information to effectively optimize a policy with respect to a latent reward function. Prior analyzed approaches fail when the feedback vector contains the action, which significantly limits IGL’s success in many potential scenarios such as Brain-computer interface (BCI) or Human-computer interface (HCI) applications. We address this by creating an algorithm and analysis which allows IGL to work even when the feedback vector contains the action, encoded in any fashion. We provide theoretical guarantees and large-scale experiments based on supervised datasets to demonstrate the effectiveness of the new approach.

## Causal Discovery in Linear Latent Variable Models Subject to Measurement Error

- Yuqin Yang · AmirEmad Ghassami · Mohamed Nafea · Negar Kiyavash · Kun Zhang · Ilya Shpitser
- abstract@[open-review](#): We focus on causal discovery in the presence of measurement error in linear systems where the mixing matrix, i.e., the matrix indicating the independent exogenous noise terms pertaining to the observed variables, is identified up to permutation and scaling of the columns. We demonstrate a somewhat surprising connection between this problem and causal discovery in the presence of unobserved parentless causes, in the sense that there is a mapping, given by the mixing matrix, between the underlying models to be inferred in these problems. Consequently, any identifiability result based on the mixing matrix for one model translates to an identifiability result for the other model. We characterize to what extent the causal models can be identified under a two-part faithfulness assumption. Under only the first part of the assumption (corresponding to the conventional definition of faithfulness), the structure can be learned up to the causal ordering among an ordered grouping of the variables but not all the edges across the groups can be identified. We further show that if both parts of the faithfulness assumption are imposed, the structure can be learned up to a more refined ordered grouping. As a result of this refinement, for the latent variable model with unobserved parentless causes, the structure can be identified. Based on our theoretical results, we propose causal structure learning methods for both models, and evaluate their performance on synthetic data.

## Thinned random measures for sparse graphs with overlapping communities

- Federica Zoe Ricci · Michele Guindani · Erik Sudderth
- abstract@[open-review](#): Network models for exchangeable arrays, including most stochastic block models, generate dense graphs with a limited ability to capture many characteristics of real-world social and biological networks. A class of models based on completely random measures like the generalized gamma process (GGP) have recently addressed some of these limitations. We propose a framework for thinning edges from realizations of GGP random graphs that models observed links via nodes' overall propensity to interact, as well as the similarity of node memberships within a large set of latent communities. Our formulation allows us to learn the number of communities from data, and enables efficient Monte Carlo methods that scale linearly with the number of observed edges, and thus (unlike dense block models) sub-quadratically with the number of entities or nodes. We compare to alternative models for both dense and sparse networks, and demonstrate effective recovery of latent community structure for real-world networks with thousands of nodes.

## Characterization of Excess Risk for Locally Strongly Convex Population Risk

- Mingyang Yi · Ruoyu Wang · Zhi-Ming Ma
- abstract@[open-review](#): We establish upper bounds for the expected excess risk of models trained by proper iterative algorithms which approximate the local minima. Unlike the results built upon the strong globally strongly convexity or global growth conditions e.g., PL-inequality, we only require the population risk to be  $\backslash$ emph{locally} strongly convex around its local minima. Concretely, our bound under convex problems is of order  $\tilde{\mathcal{O}}(1/n)$ . For non-convex problems with  $d$  model parameters such that  $d/n$  is smaller than a threshold independent of  $n$ , the order of  $\tilde{\mathcal{O}}(1/n)$  can be maintained if the empirical risk has no spurious local minima with high probability. Moreover, the bound for non-convex problem becomes  $\tilde{\mathcal{O}}(1/\sqrt{n})$  without such assumption. Our results are derived via algorithmic stability and characterization of the empirical risk's landscape. Compared with the existing algorithmic stability based results, our bounds are dimensional insensitive and without restrictions on the algorithm's implementation, learning rate, and the number of iterations. Our bounds underscore that with locally strongly convex population risk, the models trained by any proper iterative algorithm can generalize well, even for non-convex problems, and  $d$  is large.

## Learning to Attack Federated Learning: A Model-based Reinforcement Learning Attack Framework

- Henger Li · Xiaolin Sun · Zizhan Zheng
- abstract@[open-review](#): We propose a model-based reinforcement learning framework to derive untargeted poisoning attacks against federated learning (FL) systems. Our framework first approximates the distribution of the clients' aggregated data using model updates from the server. The learned distribution is then used to build a simulator of the FL environment, which is utilized to learn an adaptive attack policy through reinforcement learning. Our framework is capable of learning strong attacks automatically even when the server adopts a robust aggregation rule. We further derive an upper bound on the attacker's performance loss due to inaccurate distribution estimation. Experimental results on real-world datasets demonstrate that the proposed attack framework significantly outperforms state-of-the-art poisoning attacks. This indicates the importance of developing adaptive defenses for FL systems.

## HyperMiner: Topic Taxonomy Mining with Hyperbolic Embedding

- Yi.shi Xu · Dongsheng Wang · Bo Chen · Ruiying Lu · Zhibin Duan · Mingyuan Zhou
- abstract@[open-review](#): Embedded topic models are able to learn interpretable topics even with large and heavy-tailed vocabularies. However, they generally hold the Euclidean embedding space assumption, leading to a basic limitation in capturing hierarchical relations. To this end, we present a novel framework that introduces hyperbolic embeddings to represent words and topics. With the tree-likeness property of hyperbolic space, the underlying semantic hierarchy among words and topics can be better exploited to mine more interpretable topics. Furthermore, due to the superiority of hyperbolic geometry in representing hierarchical data, the tree-structure knowledge can be naturally injected to guide the learning of a topic hierarchy. Therefore, we further develop a regularization term based on the contrastive learning concept to efficiently inject prior structural knowledge. Experiments on both topic taxonomy discovery and document representation tasks demonstrate the proposed framework achieves improved performance on the basis of existing embedded topic models.

## On Enforcing Better Conditioned Meta-Learning for Rapid Few-Shot Adaptation

- Markus Hiller · Mehrtash Harandi · Tom Drummond
- abstract@[open-review](#): Inspired by the concept of preconditioning, we propose a novel method to increase adaptation speed for gradient-based meta-learning methods without incurring extra parameters. We demonstrate that recasting the optimisation problem to a non-linear least-squares formulation provides a principled way to actively enforce a well-conditioned parameter space for meta-learning models based on the concepts of the condition number and local curvature. Our comprehensive evaluations show that the proposed method significantly outperforms its unconstrained counterpart especially during initial adaptation steps, while achieving comparable or better overall results on several few-shot classification tasks — creating the possibility of dynamically choosing the number of adaptation steps at inference time.

## Decoupled Context Processing for Context Augmented Language Modeling

- Zonglin Li · Ruiqi Guo · Sanjiv Kumar
- abstract@[open-review](#): Language models can be augmented with context retriever to incorporate knowledge from large external databases. By leveraging retrieved context, the neural network does not have to memorize the massive amount of world knowledge within its internal parameters, leading to better parameter efficiency, interpretability and modularity. In this paper we examined a simple yet effective architecture for incorporating external context into language models based on decoupled  $\backslash$ texttt{Encoder-Decoder} $\backslash$  architecture. We showed that such a simple architecture achieves competitive results on auto-regressive language modeling and open domain question answering tasks. We also analyzed the behavior of the proposed model which performs grounded context transfer. Finally we discussed the computational implications of such retrieval augmented models.

## Cluster and Aggregate: Face Recognition with Large Probe Set

- Minchul Kim · Feng Liu · Anil K Jain · Xiaoming Liu
- abstract@[open-review](#): Feature fusion plays a crucial role in unconstrained face recognition where inputs (probes) comprise of a set of  $N$  low quality images whose individual qualities vary. Advances in attention and recurrent modules have led to feature fusion that can model the relationship among the images in the input set. However, attention mechanisms cannot scale to large  $N$  due to their quadratic complexity and recurrent modules suffer from input order sensitivity. We propose a two-stage feature fusion paradigm, Cluster and Aggregate, that can both scale to large  $N$  and maintain the ability to perform sequential inference with order invariance. Specifically, Cluster stage is a linear assignment of  $N$  inputs to  $M$  global cluster centers, and Aggregation stage is a fusion over  $M$  clustered features. The clustered features play an integral role when the inputs are sequential as they can serve as a summarization of past features. By leveraging the order-invariance of incremental averaging operation, we design an update rule that achieves batch-order invariance, which guarantees that the contributions of early image in the sequence do not diminish as time steps increase. Experiments on IJB-B and IJB-S benchmark datasets show the superiority of the proposed two-stage paradigm in unconstrained face recognition.

## [Safety Guarantees for Neural Network Dynamic Systems via Stochastic Barrier Functions](#)

- Rayan Mazouz · Karan Muvala · Akash Ratheesh Babu · Luca Laurenti · Morteza Lahijanian
- abstract@[open-review](#): Neural Networks (NNs) have been successfully employed to represent the state evolution of complex dynamical systems. Such models, referred to as NN dynamic models (NNDMs), use iterative noisy predictions of NN to estimate a distribution of system trajectories over time. Despite their accuracy, safety analysis of NNDMs is known to be a challenging problem and remains largely unexplored. To address this issue, in this paper, we introduce a method of providing safety guarantees for NNDMs. Our approach is based on stochastic barrier functions, whose relation with safety are analogous to that of Lyapunov functions with stability. We first show a method of synthesizing stochastic barrier functions for NNDMs via a convex optimization problem, which in turn provides a lower bound on the system's safety probability. A key step in our method is the employment of the recent convex approximation results for NNs to find piece-wise linear bounds, which allow the formulation of the barrier function synthesis problem as a sum-of-squares optimization program. If the obtained safety probability is above the desired threshold, the system is certified. Otherwise, we introduce a method of generating controls for the system that robustly minimize the unsafety probability in a minimally-invasive manner. We exploit the convexity property of the barrier function to formulate the optimal control synthesis problem as a linear program. Experimental results illustrate the efficacy of the method. Namely, they show that the method can scale to multi-dimensional NNDMs with multiple layers and hundreds of neurons per layer, and that the controller can significantly improve the safety probability.

## [Online Frank-Wolfe with Arbitrary Delays](#)

- Yuanyu Wan · Wei-Wei Tu · Lijun Zhang
- abstract@[open-review](#): The online Frank-Wolfe (OFW) method has gained much popularity for online convex optimization due to its projection-free property. Previous studies showed that OFW can attain an  $\mathcal{O}(T^{3/4})$  regret bound for convex losses and an  $\mathcal{O}(T^{2/3})$  regret bound for strongly convex losses. However, they assumed that each gradient queried by OFW is revealed immediately, which may not hold in practice and limits the application of OFW. To address this limitation, we proposed a delayed variant of OFW, which allows gradients to be delayed by arbitrary rounds. The main idea is to perform an update similar to OFW after receiving any delayed gradient, and play the latest decision for each round. Despite its simplicity, we prove that our delayed variant of OFW is able to achieve an  $\mathcal{O}(T^{3/4} + dT^{1/4})$  regret bound for convex losses and an  $\mathcal{O}(T^{2/3} + d\log T)$  regret bound for strongly convex losses, where  $d$  is the maximum delay. This is somewhat surprising since under a relatively large amount of delay (e.g.,  $d = \mathcal{O}(\sqrt{T})$  for convex losses and  $d = \mathcal{O}(T^{2/3}/\log T)$  for strongly convex losses), the delayed variant of OFW enjoys the same regret bound like that of the original OFW.

## [Planning for Sample Efficient Imitation Learning](#)

- Zhao-Heng Yin · Weirui Ye · Qifeng Chen · Yang Gao
- abstract@[open-review](#): Imitation learning is a class of promising policy learning algorithms that is free from many practical issues with reinforcement learning, such as the reward design issue and the exploration hardness. However, the current imitation algorithm struggles to achieve both high performance and high in-environment sample efficiency simultaneously. Behavioral Cloning~(BC) does not need in-environment interactions, but it suffers from the covariate shift problem which harms its performance. Adversarial Imitation Learning~(AIL) turns imitation learning into a distribution matching problem. It can achieve better performance on some tasks but it requires a large number of in-environment interactions. Inspired by the recent success of EfficientZero in RL, we propose EfficientImitate~(EI), a planning-based imitation learning method that can achieve high in-environment sample efficiency and performance simultaneously. Our algorithmic contribution in this paper is two-fold. First, we extend AIL into the MCTS-based RL. Second, we show the seemingly incompatible two classes of imitation algorithms (BC and AIL) can be naturally unified under our framework, enjoying the benefits of both. We benchmark our method not only on the state-based DeepMind Control Suite, but also on the image version which many previous works find highly challenging. Experimental results show that EI achieves state-of-the-art results in performance and sample efficiency. EI shows over 4x gain in performance in the limited sample setting on state-based and image-based tasks and can solve challenging problems like Humanoid, where previous methods fail with small amount of interactions.

## [Improving Neural Ordinary Differential Equations with Nesterov's Accelerated Gradient Method](#)

- Ho Huu Nghia Nguyen · Tan Nguyen · Huyen Vo · Stanley Osher · Thieu Vo
- abstract@[open-review](#): We propose the Nesterov neural ordinary differential equations (NesterovNODEs), whose layers solve the second-order ordinary differential equations (ODEs) limit of Nesterov's accelerated gradient (NAG) method, and a generalization called GNesterovNODEs. Taking the advantage of the convergence rate  $\mathcal{O}(1/k^2)$  of the NAG scheme, GNesterovNODEs speed up training and inference by reducing the number of function evaluations (NFEs) needed to solve the ODEs. We also prove that the adjoint state of a GNesterovNODEs also satisfies a GNesterovNODEs, thus accelerating both forward and backward ODE solvers and allowing the model to be scaled up for large-scale tasks. We empirically corroborate the advantage of GNesterovNODEs on a wide range of practical applications, including point cloud separation, image classification, and sequence modeling. Compared to NODEs, GNesterovNODEs require a significantly smaller number of NFEs while achieving better accuracy across our experiments.

## [Learning Structure from the Ground up---Hierarchical Representation Learning by Chunking](#)

- Shuchen Wu · Noemi Elteto · Ishita Dasgupta · Eric Schulz
- abstract@[open-review](#): From learning to play the piano to speaking a new language, reusing and recombining previously acquired representations enables us to master complex skills and easily adapt to new environments. Inspired by the Gestalt principle of grouping by proximity and theories of chunking in cognitive science, we propose a hierarchical chunking model (HCM). HCM learns representations from non-i.i.d. sequential data from the ground up by first discovering the minimal atomic sequential units as chunks. As learning progresses, a hierarchy of chunk representations is acquired by chunking previously learned representations into more complex representations guided by sequential dependence. We provide learning guarantees on an idealized version of HCM, and demonstrate that HCM learns meaningful and interpretable representations in a human-like fashion. Our model can be extended to learn visual, temporal, and visual-temporal chunks. The interpretability of the learned chunks can be used to assess transfer or interference when the environment changes. Finally, in an fMRI dataset, we demonstrate that HCM learns interpretable chunks of functional coactivation regions and hierarchical modular and sub-modular structures confirmed by the neuroscientific literature. Taken together, our results show how cognitive science in general and theories of chunking in particular can inform novel and more interpretable approaches to representation learning.

## [Are Defenses for Graph Neural Networks Robust?](#)

- Felix Mujkanovic · Simon Geisler · Aleksandar Bojchevski · Stephan Günnemann
- abstract@[open-review](#): A cursory reading of the literature suggests that we made a lot of progress in designing effective adversarial defenses for Graph Neural Networks (GNNs). Yet, the standard methodology has a serious flaw --- virtually all of the defenses are evaluated against non-adaptive attacks leading to overly optimistic robustness estimates. We perform a thorough robustness analysis of 7 of the most popular defenses spanning the entire spectrum of strategies, i.e. aimed at improving the graph, the architecture, or the training. The results are sobering --- most defenses show no or only marginal improvement compared to an undefended baseline. We advocate using custom adaptive attacks as a gold standard and we outline the lessons we learned from successfully designing such attacks. Moreover, our diverse collection of perturbed graphs forms a (black-box) unit test offering a first glance at a model's robustness.

## [Giving Feedback on Interactive Student Programs with Meta-Exploration](#)

- Evan Liu · Moritz Stephan · Allen Nie · Chris Piech · Emma Brunskill · Chelsea Finn
- abstract@[open-review](#): Creating interactive software, such as websites or games, is a particularly engaging way to learn computer science. However, teaching and giving feedback on such software is hard — standard approaches require instructors to hand grade student-implemented interactive programs. As a result, online platforms that serve millions, like Code.org, are unable to provide any feedback on assignments for implementing interactive programs, which critically hinders students' ability to learn. Recent work proposes to train reinforcement learning agents to interact with a student's program, aiming to explore states indicative of errors. However, this approach only provides binary feedback of whether a program is correct or not, while students require finer-grained feedback on the specific errors in their programs to understand their mistakes. In this work, we show that exploring to discover errors can be cast as a meta-exploration problem. This enables us to construct a principled objective for discovering errors and an algorithm for optimizing this objective, which provides fine-grained feedback. We evaluate our approach on a set of 700K real anonymized student programs from a Code.org interactive assignment. Our approach provides feedback with 94.3% accuracy, improving over existing approaches by over 17.7% and coming within 1.5% of human-level accuracy.

## [Effective Adaptation in Multi-Task Co-Training for Unified Autonomous Driving](#)

- Xiwen Liang · Yangxin Wu · Jianhua Han · Hang Xu · Chunjing XU · Xiaodan Liang
- abstract@[open-review](#): Aiming towards a holistic understanding of multiple downstream tasks simultaneously, there is a need for extracting features with better transferability. Though many latest self-supervised pre-training methods have achieved impressive performance on various vision tasks under the prevailing pretrain-finetune paradigm, their generalization capacity to multi-task learning scenarios is yet to be explored. In this paper, we extensively investigate the transfer performance of various types of self-supervised methods, e.g., MoCo and SimCLR, on three downstream tasks, including semantic segmentation, drivable area segmentation, and traffic object detection, on the large-scale driving dataset BDD100K. We surprisingly find that their performances are sub-optimal or even lag far behind the single-task baseline, which may be due to the distinctions of training objectives and architectural design lied in the pretrain-finetune paradigm. To overcome this dilemma as well as avoid redesigning the resource-intensive pre-training stage, we propose a simple yet effective pretrain-adapt-finetune paradigm for general multi-task training, where the off-the-shelf pretrained models can be effectively adapted without increasing the training overhead. During the adapt stage, we utilize learnable multi-scale adapters to dynamically adjust the pretrained model weights supervised by multi-task objectives while leaving the pretrained knowledge untouched. Furthermore, we regard the vision-language pre-training model CLIP as a strong complement to the pretrain-adapt-finetune paradigm and propose a novel adapter named LV-Adapter, which incorporates language priors in the multi-task model via task-specific prompting and alignment between visual and textual features. Our experiments demonstrate that the adapt stage significantly improves the overall performance of those off-the-shelf pretrained models and the contextual features generated by LV-Adapter are of general benefits for downstream tasks.

## [SPD domain-specific batch normalization to crack interpretable unsupervised domain adaptation in EEG](#)

- Reinmar Kobler · Jun-ichiro Hirayama · Qibin Zhao · Motoaki Kawanabe
- abstract@[open-review](#): Electroencephalography (EEG) provides access to neuronal dynamics non-invasively with millisecond resolution, rendering it a viable method in neuroscience and healthcare. However, its utility is limited as current EEG technology does not generalize well across domains (i.e., sessions and subjects) without expensive supervised re-calibration. Contemporary methods cast this transfer learning (TL) problem as a multi-source/-target unsupervised domain adaptation (UDA) problem and address it with deep learning or shallow, Riemannian geometry aware alignment methods. Both directions have, so far, failed to consistently close the performance gap to state-of-the-art domain-specific methods based on tangent space mapping (TSM) on the symmetric, positive definite (SPD) manifold. Here, we propose a machine learning framework that enables, for the first time, learning domain-invariant TSM models in an end-to-end fashion. To achieve this, we propose a new building block for geometric deep learning, which we denote SPD domain-specific momentum batch normalization (SPDDSMBN). A SPDDSMBN layer can transform domain-specific SPD inputs into domain-invariant SPD outputs, and can be readily applied to multi-source/-target and online UDA scenarios. In extensive experiments with 6 diverse EEG brain-computer interface (BCI) datasets, we obtain state-of-the-art performance in inter-session and -subject TL with a simple, intrinsically interpretable network architecture, which we denote TSMNet.

## [Geodesic Self-Attention for 3D Point Clouds](#)

- Zhengyu Li · Zihao Xu · Xihao Wang · XUAN TANG · Mingsong Chen · Hui Yu · xian wei
- abstract@[open-review](#): Due to the outstanding competence in capturing long-range relationships, the Self-attention mechanism has achieved remarkable progress in point cloud tasks. Nevertheless, point cloud object often has complex non-Euclidean spatial structures, with the behavior changing dynamically and unpredictably. The current self-attention module highly relies on the dot product multiplication in Euclidean space, which cannot capture internal non-Euclidean structures of point cloud objects, especially the long relationships along the curve of the manifold surface of point cloud object. To address this problem, in this paper, we introduce the metric on the Riemannian manifold to capture the long-range geometrical dependencies of point cloud objects to replace traditional self-attention modules, namely, the Geodesic Self-attention (GSA) module. Our approach achieves state-of-the-art performance on object classification, few-shot learning and part segmentation benchmarks.

## [Counterfactual Neural Temporal Point Process for Misinformation Impact Estimation on Social Media](#)

- Yizhou Zhang · Defu Cao · Yan Liu
- abstract@[open-review](#): Recent years have witnessed the rise of misinformation campaign which spread specific narratives on social media to manipulate public opinions on different areas, such as politics and healthcare. Consequently, an effective and efficient automatic methodology to estimate the impact of the misinformation on user beliefs and activities. However, existing works on misinformation impact estimation either rely on small-scale psychological experiments or can only discover the correlation between user behaviour and misinformation. To address these issues, in this paper, we build up a causal framework that model the causal effect of misinformation from the perspective of temporal point process. To adapt the large-scale data, we design an efficient yet precise way to estimate the ITE via neural temporal point process and gaussian mixture models. Extensive experiments on synthetic dataset and real-world dataset verify the effectiveness and efficiency of our model.

## [Smoothed Embeddings for Certified Few-Shot Learning](#)

- Mikhail Pautov · Olesya Kuznetsova · Nurislam Tursynbek · Aleksandr Petushko · Ivan Oseledets
- abstract@[open-review](#): Randomized smoothing is considered to be the state-of-the-art provable defense against adversarial perturbations. However, it heavily exploits the fact that classifiers map input objects to class probabilities and do not focus on the ones that learn a metric space in which classification is performed by computing distances to embeddings of classes prototypes. In this work, we extend randomized smoothing to few-shot learning models that map inputs to normalized embeddings. We provide analysis of Lipschitz continuity of such models and derive robustness certificate against  $\ell_2$ -bounded perturbations that may be useful in few-shot learning scenarios. Our theoretical results are confirmed by experiments on different datasets.

## [Training Subset Selection for Weak Supervision](#)

- Hunter Lang · Aravindan Vijayaraghavan · David Sontag
- abstract@[open-review](#): Existing weak supervision approaches use all the data covered by weak signals to train a classifier. We show both theoretically and empirically that this is not always optimal. Intuitively, there is a tradeoff between the amount of weakly-labeled data and the precision of the weak labels. We explore this tradeoff by combining pretrained data representations with the cut statistic to select (hopefully) high-quality subsets of the weakly-labeled training data. Subset selection applies to any label model and classifier and is very simple to plug in to existing weak supervision pipelines, requiring just a few lines of code. We show our subset selection method improves the performance of weak supervision for a wide range of label models, classifiers, and datasets. Using less weakly-labeled data improves the accuracy of weak supervision pipelines by up to 19% (absolute) on benchmark tasks.

## [Fair Wrapping for Black-box Predictions](#)

- Alexander Soen · Ibrahim Alabdulmohsin · Sanmi Koyejo · Yishay Mansour · Nyalleng Moorosi · Richard Nock · Ke Sun · Lexing Xie
- abstract@[open-review](#): We introduce a new family of techniques to post-process ("wrap") a black-box classifier in order to reduce its bias. Our technique builds on the recent analysis of improper loss functions whose optimization can correct any twist in prediction, unfairness being treated as a twist. In the post-processing, we learn a wrapper function which we define as an  $\alpha$ -tree, which modifies the prediction. We provide two generic boosting algorithms to learn  $\alpha$ -trees. We show that our modification has appealing properties in terms of composition of  $\alpha$ -trees, generalization, interpretability, and KL divergence between modified and original predictions. We exemplify the use of our technique in three fairness notions: conditional value at risk, equality of opportunity, and statistical parity; and provide experiments on several readily available datasets.

## [Learning in Observable POMDPs, without Computationally Intractable Oracles](#)

- Noah Golowich · Ankur Moitra · Dhruv Rohatgi
- abstract@[open-review](#): Much of reinforcement learning theory is built on top of oracles that are computationally hard to implement. Specifically for learning near-optimal policies in Partially Observable Markov Decision Processes (POMDPs), existing algorithms either need to make strong assumptions about the model dynamics (e.g. deterministic transitions) or assume access to an oracle for solving a hard optimistic planning or estimation problem as a subroutine. In this work we develop the first oracle-free learning algorithm for POMDPs under reasonable assumptions. Specifically, we give a quasipolynomial-time end-to-end algorithm for learning in "observable" POMDPs, where observability is the assumption that well-separated distributions over states induce well-separated distributions over observations. Our techniques circumvent the more traditional approach of using the principle of optimism under uncertainty to promote exploration, and instead give a novel application of barycentric spanners to constructing policy covers.

## [Robust Generalized Method of Moments: A Finite Sample Viewpoint](#)

- Dhruv Rohatgi · Vasilis Syrgkanis
- abstract@[open-review](#): For many inference problems in statistics and econometrics, the unknown parameter is identified by a set of moment conditions. A generic method of solving moment conditions is the Generalized Method of Moments (GMM). However, classical GMM estimation is potentially very sensitive to outliers. Robustified GMM estimators have been developed in the past, but suffer from several drawbacks: computational intractability, poor dimension-dependence, and no quantitative recovery guarantees in the presence of a constant fraction of outliers. In this work, we develop the first computationally efficient GMM estimator (under intuitive assumptions) that can tolerate a constant  $\epsilon$  fraction of adversarially corrupted samples, and that has an  $\ell_2$  recovery guarantee of  $O(\sqrt{\epsilon})$ . To achieve this, we draw upon and extend a recent line of work on algorithmic robust statistics for related but simpler problems such as mean estimation, linear regression and stochastic optimization. As a special case, we apply our algorithm to instrumental variables linear regression with heterogeneous treatment effects, and experimentally demonstrate that it can tolerate as much as 10% -- 15% corruption, significantly improving upon baseline methods.

## [COLD Decoding: Energy-based Constrained Text Generation with Langevin Dynamics](#)

- Lianhui Qin · Sean Welleck · Daniel Khashabi · Yejin Choi
- abstract@[open-review](#): Many applications of text generation require incorporating different constraints to control the semantics or style of generated text. These constraints can be hard (e.g., ensuring certain keywords are included in the output) and soft (e.g., contextualizing the output with the left- or right-hand context). In this paper, we present Energy-based Constrained Decoding with Langevin Dynamics (Cold), a decoding framework which unifies constrained generation as specifying constraints through an energy function, then performing efficient differentiable reasoning over the constraints through gradient-based sampling. Cold decoding is a flexible framework that can be applied directly to off-the-shelf left-to-right language models without the need for any task-specific fine-tuning, as demonstrated through three challenging text generation applications: lexically-constrained generation, abductive reasoning, and counterfactual reasoning. Our experiments on these constrained generation tasks point to the effectiveness of our approach, both in terms of automatic and human evaluation.

## [Learning the Structure of Large Networked Systems Obeying Conservation Laws](#)

- Anirudh Rayas · Rajasekhar Anguluri · Gautam Dasarathy
- abstract@[open-review](#): Many networked systems such as electric networks, the brain, and social networks of opinion dynamics are known to obey conservation laws. Examples of this phenomenon include the Kirchoff laws in electric networks and opinion consensus in social networks. Conservation laws in networked systems are modeled as balance equations of the form  $X = B^* Y$ , where the sparsity pattern of  $B^* \in \mathbb{R}^{p \times p}$  captures the connectivity of the network on  $p$  nodes, and  $Y, X \in \mathbb{R}^p$  are vectors of "potentials" and "injected flows" at the nodes respectively. The node potentials  $Y$  cause flows across edges which aim to balance out the potential difference, and the flows  $X$  injected at the nodes are extraneous to the network dynamics. In several practical systems, the network structure is often unknown and needs to be estimated from data to facilitate modeling, management, and control. To this end, one has access to samples of the node potentials  $Y$ , but only the statistics of the node injections  $X$ . Motivated by this important problem, we study the estimation of the sparsity structure of the matrix  $B^*$  from  $n$  samples of  $Y$  under the assumption that the node injections  $X$  follow a Gaussian distribution with a known covariance  $\Sigma_X$ . We propose a new  $\ell_1$ -regularized maximum likelihood estimator for tackling this problem in the high-dimensional regime where the size of the network may be vastly larger than the number of samples  $n$ . We show that this optimization problem is convex in the objective and admits a unique solution. Under a new mutual incoherence condition, we establish sufficient conditions on the triple  $(n, p, d)$  for which exact sparsity recovery of  $B^*$  is possible with high probability;  $d$  is the degree of the underlying graph. We also establish guarantees for the recovery of  $B^*$  in the element-wise maximum, Frobenius, and operator norms. Finally, we complement these theoretical results with experimental validation of the performance of the proposed estimator on synthetic and real-world data.

## [Implicit Neural Representations with Levels-of-Experts](#)

- Zekun Hao · Arun Mallya · Serge Belongie · Ming-Yu Liu
- abstract@[open-review](#): Coordinate-based networks, usually in the forms of MLPs, have been successfully applied to the task of predicting high-frequency but low-dimensional signals using coordinate inputs. To scale them to model large-scale signals, previous works resort to hybrid representations, combining a coordinate-based network with a grid-based representation, such as sparse voxels. However, such approaches lack a compact global latent representation in its grid, making it difficult to model a distribution of signals, which is important for generalization tasks. To address the limitation, we propose the Levels-of-Experts (LoE) framework, which is a novel coordinate-based representation consisting of an MLP with periodic, position-dependent weights arranged hierarchically. For each linear layer of the MLP, multiple candidate values of its weight matrix are tiled and replicated across

the input space, with different layers replicating at different frequencies. Based on the input, only one of the weight matrices is chosen for each layer. This greatly increases the model capacity without incurring extra computation or compromising generalization capability. We show that the new representation is an efficient and competitive drop-in replacement for a wide range of tasks, including signal fitting, novel view synthesis, and generative modeling.

## [A Closer Look at Learned Optimization: Stability, Robustness, and Inductive Biases](#)

- James Harrison · Luke Metz · Jascha Sohl-Dickstein
- abstract@[open-review](#): Learned optimizers---neural networks that are trained to act as optimizers---have the potential to dramatically accelerate training of machine learning models. However, even when meta-trained across thousands of tasks at huge computational expense, blackbox learned optimizers often struggle with stability and generalization when applied to tasks unlike those in their meta-training set. In this paper, we use tools from dynamical systems to investigate the inductive biases and stability properties of optimization algorithms, and apply the resulting insights to designing inductive biases for blackbox optimizers. Our investigation begins with a noisy quadratic model, where we characterize conditions in which optimization is stable, in terms of eigenvalues of the training dynamics. We then introduce simple modifications to a learned optimizer's architecture and meta-training procedure which lead to improved stability, and improve the optimizer's inductive bias. We apply the resulting learned optimizer to a variety of neural network training tasks, where it outperforms the current state of the art learned optimizer---at matched optimizer computational overhead---with regard to optimization performance and meta-training speed, and is capable of generalization to tasks far different from those it was meta-trained on.

## [A Causal Analysis of Harm](#)

- Sander Beckers · Hana Chockler · Joseph Halpern
- abstract@[open-review](#): As autonomous systems rapidly become ubiquitous, there is a growing need for a legal and regulatory framework to address when and how such a system harms someone. There have been several attempts within the philosophy literature to define harm, but none of them has proven capable of dealing with the many examples that have been presented, leading some to suggest that the notion of harm should be abandoned and ``replaced by more well-behaved notions''. As harm is generally something that is caused, most of these definitions have involved causality at some level. Yet surprisingly, none of them makes use of causal models and the definitions of actual causality that they can express. In this paper we formally define a qualitative notion of harm that uses causal models and is based on a well-known definition of actual causality (Halpern, 2016). The key novelty of our definition is that it is based on contrastive causation and uses a default utility to which the utility of actual outcomes is compared. We show that our definition is able to handle the examples from the literature, and illustrate its importance for reasoning about situations involving autonomous systems.

## [Sublinear Algorithms for Hierarchical Clustering](#)

- Arpit Agarwal · Sanjeev Khanna · Huan Li · Prathamesh Patil
- abstract@[open-review](#): Hierarchical clustering over graphs is a fundamental task in data mining and machine learning with applications in many domains including phylogenetics, social network analysis, and information retrieval. Specifically, we consider the recently popularized objective function for hierarchical clustering due to Dasgupta~\cite{Dasgupta16}, namely, minimum cost hierarchical partitioning. Previous algorithms for (approximately) minimizing this objective function require linear time/space complexity. In many applications the underlying graph can be massive in size making it computationally challenging to process the graph even using a linear time/space algorithm. As a result, there is a strong interest in designing algorithms that can perform global computation using only sublinear resources (space, time, and communication). The focus of this work is to study hierarchical clustering for massive graphs under three well-studied models of sublinear computation which focus on space, time, and communication, respectively, as the primary resources to optimize: (1) (dynamic) streaming model where edges are presented as a stream, (2) query model where the graph is queried using neighbor and degree queries, (3) massively parallel computation (MPC) model where the edges of the graph are partitioned over several machines connected via a communication channel. We design sublinear algorithms for hierarchical clustering in all three models above. At the heart of our algorithmic results is a view of the objective in terms of cuts in the graph, which allows us to use a relaxed notion of cut sparsifiers to do hierarchical clustering while introducing only a small distortion in the objective function. Our main algorithmic contributions are then to show how cut sparsifiers of the desired form can be efficiently constructed in the query model and the MPC model. We complement our algorithmic results by establishing nearly matching lower bounds that rule out the possibility of designing algorithms with better performance guarantees in each of these models.

## [A Communication-efficient Algorithm with Linear Convergence for Federated Minimax Learning](#)

- Zhenyu Sun · Ermin Wei
- abstract@[open-review](#): In this paper, we study a large-scale multi-agent minimax optimization problem, which models many interesting applications in statistical learning and game theory, including Generative Adversarial Networks (GANs). The overall objective is a sum of agents' private local objective functions. We first analyze an important special case, empirical minimax problem, where the overall objective approximates a true population minimax risk by statistical samples. We provide generalization bounds for learning with this objective through Rademacher complexity analysis. Then, we focus on the federated setting, where agents can perform local computation and communicate with a central server. Most existing federated minimax algorithms either require communication per iteration or lack performance guarantees with the exception of Local Stochastic Gradient Descent Ascent (SGDA), a multiple-local-update descent ascent algorithm which guarantees convergence under a diminishing stepsize. By analyzing Local SGDA under the ideal condition of no gradient noise, we show that generally it cannot guarantee exact convergence with constant stepsizes and thus suffers from slow rates of convergence. To tackle this issue, we propose FedGDA-GT, an improved Federated (Fed) Gradient Descent Ascent (GDA) method based on Gradient Tracking (GT). When local objectives are Lipschitz smooth and strongly-convex-strongly-concave, we prove that FedGDA-GT converges linearly with a constant stepsize to global  $\$epsilon$$ -approximation solution with  $\mathcal{O}(\log(1/\$epsilon))$  rounds of communication, which matches the time complexity of centralized GDA method. Finally, we numerically show that FedGDA-GT outperforms Local SGDA.

## [On Efficient Online Imitation Learning via Classification](#)

- Yichen Li · Chicheng Zhang
- abstract@[open-review](#): Imitation learning (IL) is a general learning paradigm for sequential decision-making problems. Interactive imitation learning, where learners can interactively query for expert annotations, has been shown to achieve provably superior sample efficiency guarantees compared with its offline counterpart or reinforcement learning. In this work, we study classification-based online imitation learning (abbrev. COIL) and the fundamental feasibility to design oracle-efficient regret-minimization algorithms in this setting. We make the following contributions: (1) we show that in the COIL problem, any proper online learning algorithm cannot guarantee a sublinear regret in general; (2) we propose Logger, an improper online learning algorithmic framework, that reduces COIL to online linear optimization, by utilizing a new definition of mixed policy class; (3) we design two oracle-efficient algorithms within the Logger framework that enjoy different sample and interaction round complexity tradeoffs, and show their improvements over behavior cloning; (4) we show that under standard complexity-theoretic assumptions, efficient dynamic regret minimization is infeasible in the Logger framework.

## [Subgame Solving in Adversarial Team Games](#)

- Brian Zhang · Luca Carminati · Federico Cacciamani · Gabriele Farina · Pierrick Olivieri · Nicola Gatti · Tuomas Sandholm
- abstract@[open-review](#): In adversarial team games, a team of players sequentially faces a team of adversaries. These games are the simplest setting with multiple players where cooperation and competition coexist, and it is known that the information asymmetry among the team members makes equilibrium approximation computationally hard. Although much effort has been spent designing scalable algorithms, the problem of solving large game instances is

open. In this paper, we extend the successful approach of solving huge two-player zero-sum games, where a blueprint strategy is computed offline by using an abstract version of the game and then it is refined online, that is, during a playthrough. In particular, to the best of our knowledge, our paper provides the first method for online strategy refinement via subgame solving in adversarial team games. Our method, based on the team belief DAG, generates a gadget game and then refine the blueprint strategy by using column-generation approaches in anytime fashion. If the blueprint is sparse, then our whole algorithm runs end-to-end in polynomial time given a best-response oracle; in particular, it avoids expanding the whole team belief DAG, which has exponential worst-case size. We apply our method to a standard test suite, and we empirically show the performance improvement of the strategies thanks to subgame solving.

## [Learning to Compare Nodes in Branch and Bound with Graph Neural Networks](#)

- Abdel Ghani Labassi · Didier Chetelat · Andrea Lodi
- abstract@[open-review](#): Branch-and-bound approaches in integer programming require ordering portions of the space to explore next, a problem known as node comparison. We propose a new siamese graph neural network model to tackle this problem, where the nodes are represented as bipartite graphs with attributes. Similar to prior work, we train our model to imitate a diving oracle that plunges towards the optimal solution. We evaluate our method by solving the instances in a plain framework where the nodes are explored according to their rank. On three NP-hard benchmarks chosen to be particularly primal-difficult, our approach leads to faster solving and smaller branch-and-bound trees than the default ranking function of the open-source solver SCIP, as well as competing machine learning methods. Moreover, these results generalize to instances larger than used for training.

## [Generalization for multiclass classification with overparameterized linear models](#)

- Vignesh Subramanian · Rahul Arya · Anant Sahai
- abstract@[open-review](#): Via an overparameterized linear model with Gaussian features, we provide conditions for good generalization for multiclass classification of minimum-norm interpolating solutions in an asymptotic setting where both the number of underlying features and the number of classes scale with the number of training points. The survival/contamination analysis framework for understanding the behavior of overparameterized learning problems is adapted to this setting, revealing that multiclass classification qualitatively behaves like binary classification in that, as long as there are not too many classes (made precise in the paper), it is possible to generalize well even in settings where regression tasks would not generalize. Besides various technical challenges, it turns out that the key difference from the binary classification setting is that there are relatively fewer training examples of each class in the multiclass setting as the number of classes increases, making the multiclass problem ``harder'' than the binary one.

## [Dance of SNN and ANN: Solving binding problem by combining spike timing and reconstructive attention](#)

- Hao Zheng · Luping Shi · Rong Zhao · Hui Lin
- abstract@[open-review](#): The binding problem is one of the fundamental challenges that prevent the artificial neural network (ANNs) from a compositional understanding of the world like human perception, because disentangled and distributed representations of generative factors can interfere and lead to ambiguity when complex data with multiple objects are presented. In this paper, we propose a brain-inspired unsupervised hybrid neural network (HNN) that introduces temporal binding theory originated from neuroscience into ANNs by integrating spike timing dynamics (via spiking neural networks, SNNs) with reconstructive attention (by ANNs). Spike timing provides an additional dimension for grouping, while reconstructive feedback coordinates the spikes into temporal coherent states. Through iterative interaction of ANN and SNN, the model continuously binds multiple objects at alternative synchronous firing times in the SNN coding space. The effectiveness of the model is evaluated on five artificially generated datasets of binary images. By visualization and analysis, we demonstrate that the binding is explainable, soft, flexible, and hierarchical. Notably, the model is trained on single object datasets without explicit supervision on grouping, but can successfully bind multiple objects on test datasets, showing its compositional generalization capability. Further results show its binding ability in dynamic situations.

## [Structural Knowledge Distillation for Object Detection](#)

- Philip de Rijk · Lukas Schneider · Marius Cordts · Dariu Gavrila
- abstract@[open-review](#): Knowledge Distillation (KD) is a well-known training paradigm in deep neural networks where knowledge acquired by a large teacher model is transferred to a small student. KD has proven to be an effective technique to significantly improve the student's performance for various tasks including object detection. As such, KD techniques mostly rely on guidance at the intermediate feature level, which is typically implemented by minimizing an  $\|\cdot\|_p$ -norm distance between teacher and student activations during training. In this paper, we propose a replacement for the pixel-wise independent  $\|\cdot\|_p$ -norm based on the structural similarity (SSIM). By taking into account additional contrast and structural cues, more information within intermediate feature maps can be preserved. Extensive experiments on MSCOCO demonstrate the effectiveness of our method across different training schemes and architectures. Our method adds only little computational overhead, is straightforward to implement and at the same time it significantly outperforms the standard  $\|\cdot\|_p$ -norms. Moreover, more complex state-of-the-art KD methods using attention-based sampling mechanisms are outperformed, including a +3.5 AP gain using a Faster R-CNN R-50 compared to a vanilla model.

## [JAW: Guaranteed Predictive Inference under Covariate Shift](#)

- Drew Prinster · Anqi Liu · Suchi Saria
- abstract@[open-review](#): We propose \textbf{JAWS}, a series of wrapper methods for distribution-free uncertainty quantification tasks under covariate shift, centered on the core method \textbf{JAW}, the \textbf{JA}ckknife+\textbf{W}eighted with data-dependent likelihood-ratio weights. JAWS also includes computationally efficient \textbf{A}pproximations of JAW using higher-order influence functions: \textbf{JAWA}. Theoretically, we show that JAW relaxes the jackknife+'s assumption of data exchangeability to achieve the same finite-sample coverage guarantee even under covariate shift. JAWA further approaches the JAW guarantee in the limit of either the sample size or the influence function order under common regularity assumptions. Moreover, we propose a general approach to repurposing any distribution-free uncertainty quantification method and its guarantees to the task of risk assessment: a task that estimates the probability that the true label lies within a user-specified interval. We then propose \textbf{JAW-R} and \textbf{JAWA-R} as the repurposed versions of proposed methods for \textbf{R}isk assessment. Practically, JAWS outperform the state-of-the-art predictive inference baselines in a variety of biased real world data sets for both interval-generation and risk-assessment predictive uncertainty auditing tasks.

## [Efficiently Factorizing Boolean Matrices using Proximal Gradient Descent](#)

- Sebastian Dalleiger · Jilles Vreeken
- abstract@[open-review](#): Addressing the interpretability problem of NMF on Boolean data, Boolean Matrix Factorization (BMF) uses Boolean algebra to decompose the input into low-rank Boolean factor matrices. These matrices are highly interpretable and very useful in practice, but they come at the high computational cost of solving an NP-hard combinatorial optimization problem. To reduce the computational burden, we relax BMF using a novel elastic binary regularizer, from which we derive an efficient proximal point algorithm. Through an extensive set of experiments, we demonstrate that our method works well in practice: On synthetic data, we show that our algorithm converges quickly, recovers the ground truth precisely, and estimates the true matrix rank robustly. On real-world data, we improve upon the state of the art in recall, loss, and runtime, and a case study from the medical domain confirms that our results are easily interpretable and semantically meaningful.

## The Power and Limitation of Pretraining-Finetuning for Linear Regression under Covariate Shift

- Jingfeng Wu · Difan Zou · Vladimir Braverman · Quanquan Gu · Sham Kakade
- abstract@[open-review](#): We study linear regression under covariate shift, where the marginal distribution over the input covariates differs in the source and the target domains, while the conditional distribution of the output given the input covariates is similar across the two domains. We investigate a transfer learning approach with pretraining on the source data and finetuning based on the target data (both conducted by online SGD) for this problem. We establish sharp instance-dependent excess risk upper and lower bounds for this approach. Our bounds suggest that for a large class of linear regression instances, transfer learning with  $\mathcal{O}(N^2)$  source data (and scarce or no target data) is as effective as supervised learning with  $N$  target data. In addition, we show that finetuning, even with only a small amount of target data, could drastically reduce the amount of source data required by pretraining. Our theory sheds light on the effectiveness and limitation of pretraining as well as the benefits of finetuning for tackling covariate shift problems.

## Unsupervised Reinforcement Learning with Contrastive Intrinsic Control

- Michael Laskin · Hao Liu · Xue Bin Peng · Denis Yarats · Aravind Rajeswaran · Pieter Abbeel
- abstract@[open-review](#): We introduce Contrastive Intrinsic Control (CIC), an unsupervised reinforcement learning (RL) algorithm that maximizes the mutual information between state-transitions and latent skill vectors. CIC utilizes contrastive learning between state-transitions and skills vectors to learn behaviour embeddings and maximizes the entropy of these embeddings as an intrinsic reward to encourage behavioural diversity. We evaluate our algorithm on the Unsupervised RL Benchmark (URLB) in the asymptotic state-based setting, which consists of a long reward-free pre-training phase followed by a short adaptation phase to downstream tasks with extrinsic rewards. We find that CIC improves over prior exploration algorithms in terms of adaptation efficiency to downstream tasks on state-based URLB.

## Prompt Certified Machine Unlearning with Randomized Gradient Smoothing and Quantization

- Zijie Zhang · Xin Zhao · Tianshi Che · Yang Zhou · Lingjuan Lyu
- abstract@[open-review](#): The right to be forgotten calls for efficient machine unlearning techniques that make trained machine learning models forget a cohort of data. The combination of training and unlearning operations in traditional machine unlearning methods often leads to the expensive computational cost on large-scale data. This paper presents a prompt certified machine unlearning algorithm, PCMU, which executes one-time operation of simultaneous training and unlearning in advance for a series of machine unlearning requests, without the knowledge of the removed/forgotten data. First, we establish a connection between randomized smoothing for certified robustness on classification and randomized smoothing for certified machine unlearning on gradient quantization. Second, we propose a prompt certified machine unlearning model based on randomized data smoothing and gradient quantization. We theoretically derive the certified radius  $R$  regarding the data change before and after data removals and the certified budget of data removals about  $R$ . Last but not least, we present another practical framework of randomized gradient smoothing and quantization, due to the dilemma of producing high confidence certificates in the first framework. We theoretically demonstrate the certified radius  $R'$  regarding the gradient change, the correlation between two types of certified radii, and the certified budget of data removals about  $R'$ .

## Cryptographic Hardness of Learning Halfspaces with Massart Noise

- Ilias Diakonikolas · Daniel Kane · Pasin Manurangsi · Lisheng Ren
- abstract@[open-review](#): We study the problem of PAC learning halfspaces in the presence of Massart noise. In this model, we are given i.i.d. labeled examples  $(\mathbf{x}, y) \in \mathbb{R}^N \times \{\pm 1\}$ , where the distribution of  $\mathbf{x}$  is arbitrary and the label  $y$  is a Massart corruption of  $f(\mathbf{x})$ , for an unknown halfspace  $f: \mathbb{R}^N \rightarrow \{\pm 1\}$ , with flipping probability  $\eta(\mathbf{x}) \leq \eta < 1/2$ . The goal is to compute a hypothesis with small 0-1 error. We give a reduction of the Learning with Errors (LWE) problem to the Massart halfspace learning problem. Assuming the (widely believed) subexponential-time hardness of the LWE problem, we show that no polynomial-time learner for Massart halfspaces can achieve error better than  $\Omega(\eta)$ , even if  $\text{OPT} = \mathbf{E}_{\mathbf{x}}[\eta(\mathbf{x})] = 2^{-\log c}(N)$ , for any universal constant  $c \in (0, 1)$ . Prior work provided qualitatively similar evidence of hardness in the Statistical Query model. Our hardness result shows that known learning algorithms for Massart halfspaces are essentially best possible.

## Hyperparameter Sensitivity in Deep Outlier Detection: Analysis and a Scalable Hyper-Ensemble Solution

- Xueying Ding · Lingxiao Zhao · Leman Akoglu
- abstract@[open-review](#): Outlier detection (OD) literature exhibits numerous algorithms as it applies to diverse domains. However, given a new detection task, it is unclear how to choose an algorithm to use, nor how to set its hyperparameter(s) (HPs) in unsupervised settings. HP tuning is an ever-growing problem with the arrival of many new detectors based on deep learning. While they have appealing properties such as task-driven representation learning and end-to-end optimization, deep models come with a long list of HPs. Surprisingly, the issue of model selection in the outlier mining literature has been ‘the elephant in the room’; a significant factor in unlocking the utmost potential of deep methods, yet little said or done to systematically tackle the issue. In the first part of this paper, we conduct the first large-scale analysis on the HP sensitivity of deep OD methods, and through more than 35,000 trained models, quantitatively demonstrate that model selection is inevitable. Next, we design a HP-robust and scalable deep hyper-ensemble model called ROBOD that assembles models with varying HP configurations, bypassing the choice paralysis. Importantly, we introduce novel strategies to speed up ensemble training, such as parameter sharing, batch/simultaneous training, and data subsampling, that allow us to train fewer models with fewer parameters. Extensive experiments on both image and tabular datasets show that ROBOD achieves and retains robust, state-of-the-art detection performance as compared to its modern counterparts, while taking only 2–10% of the time by the naive hyper-ensemble with independent training.

## Learning to Follow Instructions in Text-Based Games

- Mathieu Tuli · Andrew Li · Pashootan Vaezipoor · Toryn Klassen · Scott Sanner · Sheila McIlraith
- abstract@[open-review](#): Text-based games present a unique class of sequential decision making problem in which agents interact with a partially observable, simulated environment via actions and observations conveyed through natural language. Such observations typically include instructions that, in a reinforcement learning (RL) setting, can directly or indirectly guide a player towards completing reward-worthy tasks. In this work, we study the ability of RL agents to follow such instructions. We conduct experiments that show that the performance of state-of-the-art text-based game agents is largely unaffected by the presence or absence of such instructions, and that these agents are typically unable to execute tasks to completion. To further study and address the task of instruction following, we equip RL agents with an internal structured representation of natural language instructions in the form of Linear Temporal Logic (LTL), a formal language that is increasingly used for temporally extended reward specification in RL. Our framework both supports and highlights the benefit of understanding the temporal semantics of instructions and in measuring progress towards achievement of such a temporally extended behaviour. Experiments demonstrate the superior performance of our approach.

## The Importance of Baselines in Policy Gradients

- Jincheng Mei · Wesley Chung · Valentin Thomas · Bo Dai · Csaba Szepesvari · Dale Schuurmans
- abstract@[open-review](#): We study the effect of baselines in on-policy stochastic policy gradient optimization, and close the gap between the theory and practice of direct policy optimization methods. Our first contribution is to show that the  $\text{state value}$  baseline allows on-policy stochastic

\emph{natural} policy gradient (NPG) to converge to an optimal policy at an  $O(1/t)$  rate, which was not previously known. The analysis relies on two novel findings: the expected progress of the NPG update satisfies a stochastic version of the non-uniform Łojasiewicz (NL) inequality, and the state value baseline prevents the optimal action's probability from vanishing with probability 1\$, thus ensuring sufficient exploration. Importantly, these results provide a new understanding of the utility of a baseline in stochastic policy gradient: by showing that the variance of natural policy gradient estimates remains unbounded with or without a baseline, we establish that variance reduction \emph{cannot} explain their utility in this setting. Instead, our analysis reveals that the primary effect of the value baseline is to reduce the aggressiveness of the updates rather than their variance. That is, we show that finite variance is not necessary for almost sure convergence of stochastic NPG, yet controlling the update aggressiveness is both necessary and sufficient. Additional experimental results are provided to verify the theoretical findings.

## [OOD Link Prediction Generalization Capabilities of Message-Passing GNNs in Larger Test Graphs](#)

- Yangze Zhou · Gitta Kutyniok · Bruno Ribeiro
- abstract@[open-review](#): This work provides the first theoretical study on the ability of graph Message Passing Neural Networks (gMPNNs) ---such as Graph Neural Networks (GNNs)--- to achieve counterfactually-invariant representations for inductive out-of-distribution (OOD) link prediction tasks, where deployment (test) graph sizes are larger than training graphs. We first prove non-asymptotic bounds showing that link predictors based on permutation-equivariant (structural) node embeddings obtained by gMPNNs can converge to a random guess as test graphs get larger. We then propose a theoretically-sound gMPNN that outputs structural pairwise (2-node) embeddings and prove non-asymptotic bounds showing that, as test graphs grow, these embeddings converge to embeddings of a continuous function that retains its ability to predict links OOD. Empirical results on random graphs show agreement with our theoretical results.

## [On Feature Learning in the Presence of Spurious Correlations](#)

- Pavel Izmailov · Polina Kirichenko · Nate Gruver · Andrew Wilson
- abstract@[open-review](#): Deep learning classifiers are known to rely on spurious correlations — patterns which are semantically irrelevant but predictive of the target on the training data. In this paper we explore the quality of feature representations learned by standard empirical risk minimization (ERM) and specialized group robustness training, as well as the effect of various factors such as the architecture, pre-training strategy, regularization and others. Following recent work on Deep Feature Reweighting (DFR), we evaluate the feature representations by re-training the last layer of the model on a held-out set where the spurious correlation is broken. Through this procedure, we understand how much information about the core semantic features is contained in the learned representations. On multiple vision and NLP problems, we show that the features learned by simple ERM are highly competitive with the features learned by specialized group robustness methods targeted at reducing the effect of spurious correlations. Moreover, we show that the quality of learned feature representations is largely affected by the choice of data augmentation, model architecture and pre-training strategy. On the other hand, we find that strong regularization, and long training are generally not helpful for improving the learned feature representations. Finally, using insights from our analysis, we significantly improve upon the best results reported in the literature on the popular Waterbirds, CelebA hair color prediction and WILDS-FMOW problems, achieving 97%, 92% and 50% worst-group accuracies respectively.

## [Online Minimax Multiobjective Optimization: Multicalibration and Other Applications](#)

- Daniel Lee · Georgy Noarov · Mallesh Pai · Aaron Roth
- abstract@[open-review](#): We introduce a simple but general online learning framework in which a learner plays against an adversary in a vector-valued game that changes every round. Even though the learner's objective is not convex-concave (and so the minimax theorem does not apply), we give a simple algorithm that can compete with the setting in which the adversary must announce their action first, with optimally diminishing regret. We demonstrate the power of our framework by using it to (re)derive optimal bounds and efficient algorithms across a variety of domains, ranging from multicalibration to a large set of no-regret algorithms, to a variant of Blackwell's approachability theorem for polytopes with fast convergence rate. As a new application, we show how to ``(multi)calibrate'' an arbitrary collection of forecasters --- achieving an exponentially improved dependence on the number of models we are competing against, compared to prior work.

## [Escaping from the Barren Plateau via Gaussian Initializations in Deep Variational Quantum Circuits](#)

- Kaining Zhang · Liu Liu · Min-Hsiu Hsieh · Dacheng Tao
- abstract@[open-review](#): Variational quantum circuits have been widely employed in quantum simulation and quantum machine learning in recent years. However, quantum circuits with random structures have poor trainability due to the exponentially vanishing gradient with respect to the circuit depth and the qubit number. This result leads to a general standpoint that deep quantum circuits would not be feasible for practical tasks. In this work, we propose an initialization strategy with theoretical guarantees for the vanishing gradient problem in general deep quantum circuits. Specifically, we prove that under proper Gaussian initialized parameters, the norm of the gradient decays at most polynomially when the qubit number and the circuit depth increase. Our theoretical results hold for both the local and the global observable cases, where the latter was believed to have vanishing gradients even for very shallow circuits. Experimental results verify our theoretical findings in the quantum simulation and quantum chemistry.

## [Learning from a Sample in Online Algorithms](#)

- C.J. Argue · Anupam Gupta · Alan Frieze · Christopher Seiler
- abstract@[open-review](#): We consider three central problems in optimization: the restricted assignment load-balancing problem, the Steiner tree network design problem, and facility location clustering. We consider the online setting, where the input arrives over time, and irrevocable decisions must be made without knowledge of the future. For all these problems, any online algorithm must incur a cost that is approximately  $\log |I|$  times the optimal cost in the worst-case, where  $|I|$  is the length of the input. But can we go beyond the worst-case? In this work we give algorithms that perform substantially better when a  $p$ -fraction of the input is given as a sample: the algorithm uses this sample to \emph{learn} a good strategy to use for the rest of the input.

## [Not All Bits have Equal Value: Heterogeneous Weight Precisions via Trainable Noise Tensors](#)

- Pedro Savarese · Xin Yuan · Yanjing Li · Michael Maire
- abstract@[open-review](#): We study the problem of training deep networks while enforcing quantization and precision constraints to its parameters, a setting which can reduce energy consumption and inference time of deployed models. Unlike previous works, we propose a method that assigns different precisions (number of bits) to weights in a neural network, yielding a heterogeneous allocation of bits across parameters. Our method is derived from a novel framework, where the intractability of optimizing discrete precisions is approximated by training per-parameter noise magnitudes. Empirical evaluations show that our approach is capable of finding highly heterogeneous precision assignments for CNNs trained on CIFAR and ImageNet, improving upon the previous state-of-the-art and offering a theoretical foundation for the design of new quantization methods.

## [When Combinatorial Thompson Sampling meets Approximation Regret](#)

- Pierre Perrault
- abstract@[open-review](#): We study the behavior of the Combinatorial Thompson Sampling policy (CTS) for combinatorial multi-armed bandit problems (CMAB), within an approximation regret setting. Although CTS has attracted a lot of interest, it has a drawback that other usual CMAB policies do not

have when considering non-exact oracles: for some oracles, CTS has a poor approximation regret (scaling linearly with the time horizon  $\$T\$$ ) [Wang and Chen, 2018]. A study is then necessary to discriminate the oracles on which CTS could learn. This study was started by Kong et al. [2021]: they gave the first approximation regret analysis of CTS for the greedy oracle, obtaining an upper bound of order  $\$\\mathcal{O}\\{\left(\\log(T)\\Delta^2\\right)\\}$ , where  $\$\\Delta\$$  is some minimal reward gap. In this paper, our objective is to push this study further than the simple case of the greedy oracle. We provide the first  $\$\\mathcal{O}\\{\left(\\log(T)\\Delta\\right)\\}$  approximation regret upper bound for CTS, obtained under a specific condition on the approximation oracle, allowing a reduction to the exact oracle analysis. We thus term this condition Reduce2Exact, and observe that it is satisfied in many concrete examples. In particular, it can be extended to the probabilistically triggered arms setting, thus capturing even more problems, such as online influence maximization.

## [Learning Audio-Visual Dynamics Using Scene Graphs](#)

- Moitreya Chatterjee · Narendra Ahuja · Anoop Cherian
- abstract@[open-review](#): There exists an unequivocal distinction between the sound produced by a static agent and that produced by a moving one, especially when the agent moves towards or away from the microphone. In this paper, we propose to use this connection between audio and visual dynamics for solving two challenging tasks simultaneously, namely: (i) separating audio sources from a mixture using visual cues, and (ii) predicting the 3D visual motion of a sounding source only using its separated audio. Towards this end, we present Audio Separator and Motion Predictor (ASMP) - a deep learning framework that leverages the 3D structure of the scene and the motion of sound sources for better audio source separation. At the heart of ASMP is a pseudo-3D scene graph capturing various objects in the video and their 3D spatial proximities. This graph is constructed by registering together 2.5D monocular depth predictions from the 2D video frames and associating the 2.5D scene regions with the outputs of an object detector applied on those frames. The audio separation task is then modeled, as a joint problem of: (i) recursively segmenting the pseudo-3D scene graph into several sub-graphs, with each associated with a constituent sound of the mixed input audio, and (ii) predicting the 3D motions of the corresponding sound sources from the separated audio. To empirically evaluate ASMP, we present experiments on two challenging audio-visual datasets, viz. Audio Separation in the Wild (ASIW) and Audio Visual Event (AVE). Our results demonstrate that ASMP achieves a clear improvement in source separation quality, outperforming prior works on both datasets, while estimating the direction of motion of the sound sources better than other methods.

## [CARD: Classification and Regression Diffusion Models](#)

- Xizewen Han · Huangjie Zheng · Mingyuan Zhou
- abstract@[open-review](#): Learning the distribution of a continuous or categorical response variable  $y$  given its covariates  $x$  is a fundamental problem in statistics and machine learning. Deep neural network-based supervised learning algorithms have made great progress in predicting the mean of  $y$  given  $x$ , but they are often criticized for their ability to accurately capture the uncertainty of their predictions. In this paper, we introduce classification and regression diffusion (CARD) models, which combine a denoising diffusion-based conditional generative model and a pre-trained conditional mean estimator, to accurately predict the distribution of  $y$  given  $x$ . We demonstrate the outstanding ability of CARD in conditional distribution prediction with both toy examples and real-world datasets, the experimental results on which show that CARD, in general, outperforms state-of-the-art methods, including Bayesian neural network-based one, designed for uncertainty estimation, especially when the conditional distribution of  $y$  given  $x$  is multi-modal.

## [Make Sharpness-Aware Minimization Stronger: A Sparsified Perturbation Approach](#)

- Peng Mi · Li Shen · Tianhe Ren · Yiyi Zhou · Xiaoshuai Sun · Rongrong Ji · Dacheng Tao
- abstract@[open-review](#): Deep neural networks often suffer from poor generalization caused by complex and non-convex loss landscapes. One of the popular solutions is Sharpness-Aware Minimization (SAM), which smooths the loss landscape via minimizing the maximized change of training loss when adding a perturbation to the weight. However, we find the indiscriminate perturbation of SAM on all parameters is suboptimal, which also will result in excessive computation, double the overhead of common optimizers like Stochastic Gradient Descent~(SGD). In this paper, we propose an efficient and effective training scheme coined as Sparse SAM (SSAM), which achieves sparse perturbation by a binary mask. To obtain the sparse mask, we provide two solutions which are based on Fisher information and dynamic sparse training, respectively. In addition, we theoretically prove that SSAM can converge at the same rate as SAM,  $\$O(\\log T\\sqrt{T})\$$ . Sparse SAM not only has the potential for training acceleration but also smooths the loss landscape effectively. Extensive experimental results on CIFAR10, CIFAR100, and ImageNet-1K confirm the superior efficiency of our method, and the performance is preserved or even better with a perturbation of merely 50% sparsity.

## [DropCov: A Simple yet Effective Method for Improving Deep Architectures](#)

- Qilong Wang · Mingze Gao · Zhaolin Zhang · Jiangtao Xie · Peihua Li · Qinghua Hu
- abstract@[open-review](#): Previous works show global covariance pooling (GCP) has great potential to improve deep architectures especially on visual recognition tasks, where post-normalization of GCP plays a very important role in final performance. Although several post-normalization strategies have been studied, these methods pay more close attention to effect of normalization on covariance representations rather than the whole GCP networks, and their effectiveness requires further understanding. Meanwhile, existing effective post-normalization strategies (e.g., matrix power normalization) usually suffer from high computational complexity (e.g.,  $\$O(d^3)\$$  for  $d$ -dimensional inputs). To handle above issues, this work first analyzes the effect of post-normalization from the perspective of training GCP networks. Particularly, we for the first time show that effective post-normalization can make a good trade-off between representation decorrelation and information preservation for GCP, which are crucial to alleviate over-fitting and increase representation ability of deep GCP networks, respectively}. Based on this finding, we can improve existing post-normalization methods with some small modifications, providing further support to our observation. Furthermore, this finding encourages us to propose a novel pre-normalization method for GCP (namely DropCov), which develops an adaptive channel dropout on features right before GCP, aiming to reach trade-off between representation decorrelation and information preservation in a more efficient way. Our DropCov only has a linear complexity of  $\$O(d)\$$ , while being free for inference. Extensive experiments on various benchmarks (i.e., ImageNet-1K, ImageNet-C, ImageNet-A, Stylized-ImageNet, and iNat2017) show our DropCov is superior to the counterparts in terms of efficiency and effectiveness, and provides a simple yet effective method to improve performance of deep architectures involving both deep convolutional neural networks (CNNs) and vision transformers (ViT).

## [Randomized Channel Shuffling: Minimal-Overhead Backdoor Attack Detection without Clean Datasets](#)

- Ruiqi Cai · Zhenyu Zhang · Tianlong Chen · Xiaohan Chen · Zhangyang Wang
- abstract@[open-review](#): Deep neural networks (DNNs) typically require massive data to train on, which is a hurdle for numerous practical domains. Facing the data shortfall, one viable option is to acquire domain-specific training data from external uncensored sources, such as the open web or the third-party data collectors. However, the quality of such acquired data is often not rigorously scrutinized, and one cannot rule out the risk of "poisoned" examples being included in such unreliable datasets. That casts a new "self-supervised" backdoor detection setting that was hardly explored by prior arts: in particular, for one specific collected dataset, users (1) have no prior knowledge whether it is poisoned, or on the target class/percentage of poisoned samples, and (2) have no access to a clean sample set from the same domain distribution, nor any trusted model trained on such clean data. This new backdoor threat hence incurs overlooked risks for many data-hungry, or high-stake applications. This paper reports the first pilot study on this new setting, by investigating the contrasting channel-level statistics between backdoor trigger and clean features, and consequently, how the former can be differentiated by progressive channel shuffling. The method leads to detecting the backdoor-targeted class with only a few feedforward passes, incurring minimal overheads, and demanding no clean sample nor prior knowledge. Extensive experiments are conducted with two datasets (CIFAR-10, GTSRB), two architectures (AlexNet, ResNet-20), and three strong attacks (BadNets, clean label attack, and WaNet). Results consistently endorse the effectiveness of our technique in backdoor model detection, with margins of 0.291~0.640 AUROC over the current state-of-the-art methods. Codes will be released.

## [Online Bipartite Matching with Advice: Tight Robustness-Consistency Tradeoffs for the Two-Stage Model](#)

- Billy Jin Â· Will Ma
- abstract@[open-review](#): We study the two-stage vertex-weighted online bipartite matching problem of Feng, Niazadeh, and Saberi (SODA â€˜21) in a setting where the algorithm has access to a suggested matching that is recommended in the first stage. We evaluate an algorithm by its robustness  $\$R\$$ , which is its performance relative to that of the optimal offline matching, and its consistency  $\$C\$$ , which is its performance when the advice or the prediction given is correct. We characterize for this problem the Pareto-efficient frontier between robustness and consistency, which is rare in the literature on advice-augmented algorithms, yet necessary for quantifying such an algorithm to be optimal. Specifically, we propose an algorithm that is  $\$R\$$ -robust and  $\$C\$$ -consistent for any  $(R,C)$  with  $0 \leq R \leq \frac{3}{4}$  and  $\sqrt{1-R} + \sqrt{1-C} = 1$ , and prove that no other algorithm can achieve a better tradeoff.

## [Few-Shot Fast-Adaptive Anomaly Detection](#)

- Ze Wang Â· Yipin Zhou Â· Rui Wang Â· Tsung-Yu Lin Â· Ashish Shah Â· Ser Nam Lim
- abstract@[open-review](#): The ability to detect anomaly has long been recognized as an inherent human ability, yet to date, practical AI solutions to mimic such capability have been lacking. This lack of progress can be attributed to several factors. To begin with, the distribution of ``abnormalities'' is intractable. Anything outside of a given normal population is by definition an anomaly. This explains why a large volume of work in this area has been dedicated to modeling the normal distribution of a given task followed by detecting deviations from it. This direction is however unsatisfying as it would require modeling the normal distribution of every task that comes along, which includes tedious data collection. In this paper, we report our work aiming to handle these issues. To deal with the intractability of abnormal distribution, we leverage Energy Based Model (EBM). EBMs learn to associate low energies to correct values and higher energies to incorrect values. At its core, the EBM employs Langevin Dynamics (LD) in generating these incorrect samples based on an iterative optimization procedure, alleviating the intractable problem of modeling the world of anomalies. Then, in order to avoid training an anomaly detector for every task, we utilize an adaptive sparse coding layer. Our intention is to design a plug and play feature that can be used to quickly update what is normal during inference time. Lastly, to avoid tedious data collection, this mentioned update of the sparse coding layer needs to be achievable with just a few shots. Here, we employ a meta learning scheme that simulates such a few shot setting during training. We support our findings with strong empirical evidence.

## [Deep Compression of Pre-trained Transformer Models](#)

- Naigang Wang Â· Chi-Chun (Charlie) Liu Â· Swagath Venkataramani Â· Sanchari Sen Â· Chia-Yu Chen Â· Kaoutar El Maghraoui Â· Vijayalakshmi (Viji) Srinivasan Â· Leland Chang
- abstract@[open-review](#): Pre-trained transformer models have achieved remarkable success in natural language processing (NLP) and have recently become competitive alternatives to Convolution Neural Networks (CNN) and Recurrent Neural Networks (RNN) in vision and speech tasks, respectively. Due to excellent computational efficiency and scalability, transformer models can be trained on exceedingly large amounts of data; however, model sizes can grow tremendously. As high performance, large-scale, and pre-trained transformer models become available for users to download and fine-tune for customized downstream tasks, the deployment of these models becomes challenging due to the vast amount of operations and large memory footprint. To address this challenge, we introduce methods to deeply compress pre-trained transformer models across three major application domains: NLP, speech, and vision. Specifically, we quantize transformer backbones down to 4-bit and further achieve 50% fine-grained structural sparsity on pre-trained BERT, Wav2vec2.0 and Vision Transformer (ViT) models to achieve 16x compression while maintaining model accuracy. This is achieved by identifying the critical initialization for quantization/sparsity aware fine-tuning, as well as novel techniques including quantizers with zero-preserving format and scheduled dropout. These hardware-friendly techniques need only to be applied in the fine-tuning phase for downstream tasks; hence, are especially suitable for acceleration and deployment of pre-trained transformer models.

## [Diffusion Visual Counterfactual Explanations](#)

- Maximilian Augustin Â· Valentyn Boreiko Â· Francesco Croce Â· Matthias Hein
- abstract@[open-review](#): An important tool to understand the decisions of an image classifier are Visual Counterfactual Explanations (VCEs), which are small "butrealistic" semantic changes of the image changing the classifier decision. Current approaches for the generation of VCEs are restricted to adversarially robust models and often contain non-realistic artefacts or are restricted to image classification problems with few classes. In this paper we overcome this by generating Diffusion Visual Counterfactual Explanations (DVCEs) for arbitrary ImageNet classifiers via a diffusion process. Two modifications to the diffusion process are key for our DVCEs: i) an adaptive parameterization allows us to have hyperparameters which generalize across images and models and together with distance regularization and late start of the diffusion process allow us to generate close images and ii) our cone regularization via an adversarially robust model ensures that the diffusion process does not converge to trivial non-semantic changes but instead produces realistic images of the target class which achieve high confidence by the classifier.

## [Sketching based Representations for Robust Image Classification with Provable Guarantees](#)

- Nishanth Dikkala Â· Sankeerth Rao Karingula Â· Raghu Meka Â· Jelani Nelson Â· Rina Panigrahy Â· Xin Wang
- abstract@[open-review](#): How do we provably represent images succinctly so that their essential latent attributes are correctly captured by the representation to as high level of detail as possible? While today's deep networks (such as CNNs) produce image embeddings they do not have any provable properties and seem to work in mysterious non-interpretable ways. In this work we theoretically study synthetic images that are composed of a union or intersection of several mathematically specified shapes using thresholded polynomial functions (for e.g. ellipses, rectangles). We show how to produce a succinct sketch of such an image so that the sketch â€œsmoothlyâ€ maps to the latent-coefficients producing the different shapes in the image. We prove several important properties such as: easy reconstruction of the image from the sketch, similarity preservation (similar shapes produce similar sketches), being able to index sketches so that other similar images and parts of other images can be retrieved, being able to store the sketches into a dictionary of concepts and shapes so parts of the same or different images that refer to the same shape can point to the same entry in this dictionary of common shape attributes.

## [Recursive Reinforcement Learning](#)

- Mateo Perez Â· Ernst Moritz Hahn Â· Sven Schewe Â· Fabio Somenzi Â· Ashutosh Trivedi Â· Dominik Wojtczak
- abstract@[open-review](#): Recursion is the fundamental paradigm to finitely describe potentially infinite objects. As state-of-the-art reinforcement learning (RL) algorithms cannot directly reason about recursion, they must rely on the practitioner's ingenuity in designing a suitable flat representation of the environment. The resulting manual feature constructions and approximations are cumbersome and error-prone; their lack of transparency hampers scalability. To overcome these challenges, we develop RL algorithms capable of computing optimal policies in environments described as a collection of Markov decision processes (MDPs) that can recursively invoke one another. Each constituent MDP is characterized by several entry and exit points that correspond to input and output values of these invocations. These recursive MDPs (or RMDPs) are expressively equivalent to probabilistic pushdown systems (with call-stack playing the role of the pushdown stack), and thus can model probabilistic programs with recursive procedural calls. We introduce Recursive Q-learning---a model-free RL algorithm for RMDPs---and prove that it converges for finite, single-exit and deterministic multi-exit RMDPs under mild assumptions.

## [On the Effectiveness of Lipschitz-Driven Rehearsal in Continual Learning](#)

- Lorenzo Bonicelli · Matteo Boschini · Angelo Porrello · Concetto Spampinato · SIMONE CALDERARA
- abstract@[open-review](#): Rehearsal approaches enjoy immense popularity with Continual Learning (CL) practitioners. These methods collect samples from previously encountered data distributions in a small memory buffer; subsequently, they repeatedly optimize on the latter to prevent catastrophic forgetting. This work draws attention to a hidden pitfall of this widespread practice: repeated optimization on a small pool of data inevitably leads to tight and unstable decision boundaries, which are a major hindrance to generalization. To address this issue, we propose Lipschitz-DrivEn Rehearsal (LiDER), a surrogate objective that induces smoothness in the backbone network by constraining its layer-wise Lipschitz constants w.r.t. replay examples. By means of extensive experiments, we show that applying LiDER delivers a stable performance gain to several state-of-the-art rehearsal CL methods across multiple datasets, both in the presence and absence of pre-training. Through additional ablative experiments, we highlight peculiar aspects of buffer overfitting in CL and better characterize the effect produced by LiDER.

## [Tractable Latent State Inference for Hidden Continuous-Time semi-Markov Chains](#)

- Nicolai Engelmann · Heinz Koeppl
- abstract@[open-review](#): Hidden semi-Markov Models (HSMM's) - while broadly in use - are restricted to a discrete and uniform time grid. They are thus not well suited to explain often irregularly spaced discrete event data from continuous-time phenomena. We show that non-sampling-based latent state inference used in HSMM's can be generalized to latent Continuous-Time semi-Markov Chains (CTSMC's). We formulate integro-differential forward and backward equations adjusted to the observation likelihood and introduce an exact integral equation for the Bayesian posterior marginals and a scalable Viterbi-type algorithm for posterior path estimates. All presented equations can be efficiently solved using existing numerical methods. We evaluate our approaches in latent state inference scenarios in comparison to classical HSMM's.

## [Asymptotics of \$\|\cdot\|\_2\$ Regularized Network Embeddings](#)

- Andrew Davison
- abstract@[open-review](#): A common approach to solving prediction tasks on large networks, such as node classification or link prediction, begin by learning a Euclidean embedding of the nodes of the network, from which traditional machine learning methods can then be applied. This includes methods such as DeepWalk and node2vec, which learn embeddings by optimizing stochastic losses formed over subsamples of the graph at each iteration of stochastic gradient descent. In this paper, we study the effects of adding an  $\|\cdot\|_2$  penalty of the embedding vectors to the training loss of these types of methods. We prove that, under some exchangeability assumptions on the graph, this asymptotically leads to learning a graphon with a nuclear-norm-type penalty, and give guarantees for the asymptotic distribution of the learned embedding vectors. In particular, the exact form of the penalty depends on the choice of subsampling method used as part of stochastic gradient descent. We also illustrate empirically that concatenating node covariates to  $\|\cdot\|_2$  regularized node2vec embeddings leads to comparable, when not superior, performance to methods which incorporate node covariates and the network structure in a non-linear manner..

## [NOMAD: Nonlinear Manifold Decoders for Operator Learning](#)

- Jacob Seidman · Georgios Kissas · Paris Perdikaris · George J. Pappas
- abstract@[open-review](#): Supervised learning in function spaces is an emerging area of machine learning research with applications to the prediction of complex physical systems such as fluid flows, solid mechanics, and climate modeling. By directly learning maps (operators) between infinite dimensional function spaces, these models are able to learn discretization invariant representations of target functions. A common approach is to represent such target functions as linear combinations of basis elements learned from data. However, there are simple scenarios where, even though the target functions form a low dimensional submanifold, a very large number of basis elements is needed for an accurate linear representation. Here we present NOMAD, a novel operator learning framework with a nonlinear decoder map capable of learning finite dimensional representations of nonlinear submanifolds in function spaces. We show this method is able to accurately learn low dimensional representations of solution manifolds to partial differential equations while outperforming linear models of larger size. Additionally, we compare to state-of-the-art operator learning methods on a complex fluid dynamics benchmark and achieve competitive performance with a significantly smaller model size and training cost.

## [Environment Diversification with Multi-head Neural Network for Invariant Learning](#)

- Bo-Wei Huang · Keng-Te Liao · Chang-Sheng Kao · Shou-De Lin
- abstract@[open-review](#): Neural networks are often trained with empirical risk minimization; however, it has been shown that a shift between training and testing distributions can cause unpredictable performance degradation. On this issue, a research direction, invariant learning, has been proposed to extract causal features insensitive to the distributional changes. This work proposes an invariant learning framework containing a multi-head neural network to absorb data biases. We show that this framework does not require prior knowledge about the environment or strong assumptions about the pre-train model. We also reveal that the proposed algorithm has theoretical connections to recent studies discussing properties of variant and invariant features. Finally, we demonstrate that empirically models trained with this framework are more robust against distributional shift.

## [Improving Diffusion Models for Inverse Problems using Manifold Constraints](#)

- Hyungjin Chung · Byeongsu Sim · Dohoon Ryu · Jong Chul Ye
- abstract@[open-review](#): Recently, diffusion models have been used to solve various inverse problems in an unsupervised manner with appropriate modifications to the sampling process. However, the current solvers, which recursively apply a reverse diffusion step followed by a projection-based measurement consistency step, often produce sub-optimal results. By studying the generative sampling path, here we show that current solvers throw the sample path off the data manifold, and hence the error accumulates. To address this, we propose an additional correction term inspired by the manifold constraint, which can be used synergistically with the previous solvers to make the iterations close to the manifold. The proposed manifold constraint is straightforward to implement within a few lines of code, yet boosts the performance by a surprisingly large margin. With extensive experiments, we show that our method is superior to the previous methods both theoretically and empirically, producing promising results in many applications such as image inpainting, colorization, and sparse-view computed tomography. Code available [https://github.com/HJ-harry/MCG\\_diffusion](https://github.com/HJ-harry/MCG_diffusion)

## [Active Bayesian Causal Inference](#)

- Christian Toth · Lars Lorch · Christian Knoll · Andreas Krause · Franz Pernkopf · Robert Peherz · Julius von Kägelgen
- abstract@[open-review](#): Causal discovery and causal reasoning are classically treated as separate and consecutive tasks: one first infers the causal graph, and then uses it to estimate causal effects of interventions. However, such a two-stage approach is uneconomical, especially in terms of actively collected interventional data, since the causal query of interest may not require a fully-specified causal model. From a Bayesian perspective, it is also unnatural, since a causal query (e.g., the causal graph or some causal effect) can be viewed as a latent quantity subject to posterior inference—quantities that are not of direct interest ought to be marginalized out in this process, thus contributing to our overall uncertainty. In this work, we propose Active Bayesian Causal Inference (ABCI), a fully-Bayesian active learning framework for integrated causal discovery and reasoning, i.e., for jointly inferring a posterior over causal models and queries of interest. In our approach to ABCI, we focus on the class of causally-sufficient nonlinear additive Gaussian noise models, which we model using Gaussian processes. To capture the space of causal graphs, we use a continuous latent graph representation, allowing our

approach to scale to practically relevant problem sizes. We sequentially design experiments that are maximally informative about our target causal query, collect the corresponding interventional data, update our beliefs, and repeat. Through simulations, we demonstrate that our approach is more data-efficient than existing methods that only focus on learning the full causal graph. This allows us to accurately learn downstream causal queries from fewer samples, while providing well-calibrated uncertainty estimates of the quantities of interest.

## [Meta-Reward-Net: Implicitly Differentiable Reward Learning for Preference-based Reinforcement Learning](#)

- Runze Liu · Fengshuo Bai · Yali Du · Yaodong Yang
- abstract@[open-review](#): Setting up a well-designed reward function has been challenging for many reinforcement learning applications. Preference-based reinforcement learning (PbRL) provides a new framework that avoids reward engineering by leveraging human preferences (i.e., preferring apples over oranges) as the reward signal. Therefore, improving the efficacy of data usage for preference data becomes critical. In this work, we propose Meta-Reward-Net (MRN), a data-efficient PbRL framework that incorporates bi-level optimization for both reward and policy learning. The key idea of MRN is to adopt the performance of the Q-function as the learning target. Based on this, MRN learns the Q-function and the policy in the inner level while updating the reward function adaptively according to the performance of the Q-function on the preference data in the outer level. Our experiments on robotic simulated manipulation tasks and locomotion tasks demonstrate that MRN outperforms prior methods in the case of few preference labels and significantly improves data efficiency, achieving state-of-the-art in preference-based RL. Ablation studies further demonstrate that MRN learns a more accurate Q-function compared to prior work and shows obvious advantages when only a small amount of human feedback is available. The source code and videos of this project are released at <https://sites.google.com/view/meta-reward-net>.

## [Asynchronous Actor-Critic for Multi-Agent Reinforcement Learning](#)

- Yuchen Xiao · Weihao Tan · Christopher Amato
- abstract@[open-review](#): Synchronizing decisions across multiple agents in realistic settings is problematic since it requires agents to wait for other agents to terminate and communicate about termination reliably. Ideally, agents should learn and execute asynchronously instead. Such asynchronous methods also allow temporally extended actions that can take different amounts of time based on the situation and action executed. Unfortunately, current policy gradient methods are not applicable in asynchronous settings, as they assume that agents synchronously reason about action selection at every time step. To allow asynchronous learning and decision-making, we formulate a set of asynchronous multi-agent actor-critic methods that allow agents to directly optimize asynchronous policies in three standard training paradigms: decentralized learning, centralized learning, and centralized training for decentralized execution. Empirical results (in simulation and hardware) in a variety of realistic domains demonstrate the superiority of our approaches in large multi-agent problems and validate the effectiveness of our algorithms for learning high-quality and asynchronous solutions.

## [CCCP is Frank-Wolfe in disguise](#)

- Alp Yurtsever · Suvrit Sra
- abstract@[open-review](#): This paper uncovers a simple but rather surprising connection: it shows that the well-known convex-concave procedure (CCCP) and its generalization to constrained problems are both special cases of the Frank-Wolfe (FW) method. This connection not only provides insight of deep (in our opinion) pedagogical value, but also transfers the recently discovered convergence theory of nonconvex Frank-Wolfe methods immediately to CCCP, closing a long-standing gap in its non-asymptotic convergence theory. We hope the viewpoint uncovered by this paper spurs the transfer of other advances made for FW to both CCCP and its generalizations.

## [MorphTE: Injecting Morphology in Tensorized Embeddings](#)

- Guobing Gan · Peng Zhang · Sunzhu Li · Xiuqing Lu · Benyou Wang
- abstract@[open-review](#): In the era of deep learning, word embeddings are essential when dealing with text tasks. However, storing and accessing these embeddings requires a large amount of space. This is not conducive to the deployment of these models on resource-limited devices. Combining the powerful compression capability of tensor products, we propose a word embedding compression method with morphological augmentation, Morphologically-enhanced Tensorized Embeddings (MorphTE). A word consists of one or more morphemes, the smallest units that bear meaning or have a grammatical function. MorphTE represents a word embedding as an entangled form of its morpheme vectors via the tensor product, which injects prior semantic and grammatical knowledge into the learning of embeddings. Furthermore, the dimensionality of the morpheme vector and the number of morphemes are much smaller than those of words, which greatly reduces the parameters of the word embeddings. We conduct experiments on tasks such as machine translation and question answering. Experimental results on four translation datasets of different languages show that MorphTE can compress word embedding parameters by about \$20\$ times without performance loss and significantly outperforms related embedding compression methods.

## [Kernel Multimodal Continuous Attention](#)

- Alexander Moreno · Zhenke Wu · Supriya Nagesh · Walter Dempsey · James Rehg
- abstract@[open-review](#): Attention mechanisms take an expectation of a data representation with respect to probability weights. Recently, (Martins et al. 2020, 2021) proposed continuous attention mechanisms, focusing on unimodal attention densities from the exponential and deformed exponential families: the latter has sparse support. (Farinhas et al 2021) extended this to multimodality via Gaussian mixture attention densities. In this paper, we extend this to kernel exponential families (Canu and Smola 2006) and our new sparse counterpart, kernel deformed exponential families. Theoretically, we show new existence results for both kernel exponential and deformed exponential families, and that the deformed case has similar approximation capabilities to kernel exponential families. Lacking closed form expressions for the context vector, we use numerical integration: we show exponential convergence for both kernel exponential and deformed exponential families. Experiments show that kernel continuous attention often outperforms unimodal continuous attention, and the sparse variant tends to highlight peaks of time series.

## [Stochastic Window Transformer for Image Restoration](#)

- Jie Xiao · Xueyang Fu · Feng Wu · Zheng-Jun Zha
- abstract@[open-review](#): Thanks to the strong representation ability, transformers have attained impressive results for image restoration. However, existing transformers do not carefully take into account the particularities of image restoration. Basically, image restoration requires that the ideal approach should be invariant to translation of degradation, i.e., undesirable degradation should be removed irrespective of its position within the image. Moreover, local relationships play a vital role and should be faithfully exploited for recovering clean images. Nevertheless, most of transformers have resorted to either fixed local window based attention or global attention, which unfortunately breaks the translation invariance and further causes huge loss of local relationships. To address these issues, we propose an elegant stochastic window strategy for transformers. Specifically, we introduce the window partition with stochastic shift to replace the original fixed window partition for training and elaborate the layer expectation propagation algorithm to efficiently approximate the expectation of the induced stochastic transformer for testing. The stochastic window transformer can not only enjoy powerful representation but also maintain the desired property of translation invariance and locality. Experiments validate the stochastic window strategy constantly improves performance on various image restoration tasks (image deraining, denosing, and deblurring) by significant margins.

## [Scalable and Efficient Non-adaptive Deterministic Group Testing](#)

- Dariusz Kowalski Å· Dominik Pajak
- abstract@[open-review](#): Group Testing (GT) is about learning a (hidden) subset  $K$ , of size  $k$ , of some large domain  $N$ , of size  $n \gg k$ , using a sequence of queries. A result of a query provides some information about the intersection of the query with the unknown set  $K$ . The goal is to design efficient (polynomial time) and scalable (polylogarithmic number of queries per element in  $K$ ) algorithms for constructing queries that allow to decode every hidden set  $K$  based on the results of the queries. A vast majority of the previous work focused on randomized algorithms minimizing the number of queries; however, in case of large domains  $N$ , randomization may result in a significant deviation from the expected precision of learning the set  $K$ . Others assumed unlimited computational power (existential results) or adaptiveness of queries (next query could be constructed taking into account the results of the previous queries) — the former approach is less practical due to non-efficiency, and the latter has several drawbacks including non-parallelization. To avoid all the abovementioned drawbacks, for Quantitative Group Testing (QGT) where query result is the size of its intersection with the hidden set, we present the first efficient and scalable non-adaptive deterministic algorithms for constructing queries and decoding a hidden set  $K$  from the results of the queries — these solutions do not use any randomization, adaptiveness or unlimited computational power.

## [Privacy Induces Robustness: Information-Computation Gaps and Sparse Mean Estimation](#)

- Kristian Georgiev Å· Samuel Hopkins
- abstract@[open-review](#): We establish a simple connection between robust and differentially-private algorithms: private mechanisms which perform well with very high probability are automatically robust in the sense that they retain accuracy even if a constant fraction of the samples they receive are adversarially corrupted. Since optimal mechanisms typically achieve these high success probabilities, our results imply that optimal private mechanisms for many basic statistics problems are robust. We investigate the consequences of this observation for both algorithms and computational complexity across different statistical problems. Assuming the Brennan-Bresler secret-leakage planted clique conjecture, we demonstrate a fundamental tradeoff between computational efficiency, privacy leakage, and success probability for sparse mean estimation. Private algorithms which match this tradeoff are not yet known -- we achieve that (up to polylogarithmic factors) in a polynomially-large range of parameters via the Sum-of-Squares method. To establish an information-computation gap for sparse mean estimation, we also design new (exponential-time) mechanisms using fewer samples than efficient algorithms must use. Finally, we give evidence for privacy-induced information-computation gaps for several other statistics and learning problems, including PAC learning parity functions and estimation of the mean of a multivariate Gaussian.

## [Learning Physics Constrained Dynamics Using Autoencoders](#)

- Tsung-Yen Yang Å· Justinian Rosca Å· Karthik Narasimhan Å· Peter J Ramadge
- abstract@[open-review](#): We consider the problem of estimating states (e.g., position and velocity) and physical parameters (e.g., friction, elasticity) from a sequence of observations when provided a dynamic equation that describes the behavior of the system. The dynamic equation can arise from first principles (e.g., Newtonâ€™s laws) and provide useful cues for learning, but its physical parameters are unknown. To address this problem, we propose a model that estimates states and physical parameters of the system using two main components. First, an autoencoder compresses a sequence of observations (e.g., sensor measurements, pixel images) into a sequence for the state representation that is consistent with physics by including a simulation of the dynamic equation. Second, an estimator is coupled with the autoencoder to predict the values of the physical parameters. We also theoretically and empirically show that using Fourier feature mappings improves generalization of the estimator in predicting physical parameters compared to raw state sequences. In our experiments on three visual and one sensor measurement tasks, our model imposes interpretability on latent states and achieves improved generalization performance for long-term prediction of system dynamics over state-of-the-art baselines.

## [Augmenting Online Algorithms with \$\backslash varepsilon\$ -Accurate Predictions](#)

- Anupam Gupta Å· Debmalya Panigrahi Å· Bernardo Subercaseaux Å· Kevin Sun
- abstract@[open-review](#): The growing body of work in learning-augmented online algorithms studies how online algorithms can be improved when given access to ML predictions about the future. Motivated by ML models that give a confidence parameter for their predictions, we study online algorithms with predictions that are  $\backslash varepsilon$ -accurate: namely, each prediction is correct with probability (at least)  $\backslash varepsilon$ , but can be arbitrarily inaccurate with the remaining probability. We show that even with predictions that are accurate with a small probability and arbitrarily inaccurate otherwise, we can dramatically outperform worst-case bounds for a range of classical online problems including caching, online set cover, and online facility location. Our main results are an  $O(\log(1/\backslash varepsilon))$ -competitive algorithm for caching, and a simple  $O(1/\backslash varepsilon)$ -competitive algorithm for a large family of covering problems, including set cover and facility location, with  $\backslash varepsilon$ -accurate predictions.

## [Envy-free Policy Teaching to Multiple Agents](#)

- Jiarui Gan Å· R Majumdar Å· Adish Singla Å· Goran Radanovic
- abstract@[open-review](#): We study envy-free policy teaching. A number of agents independently explore a common Markov decision process (MDP), but each with their own reward function and discounting rate. A teacher wants to teach a target policy to the diverse group of agents, by way of modifying the agents' reward functions, providing additional bonus to certain behaviors or penalizing others. These reward modifications are personalized for each agent. An important question in this setting concerns how a teaching program can be designed so that the agents think that they are treated fairly. We adopt the fairness notion of envy-freeness (EF) to formalize this question and define three different EF notions, each imposing stronger requirements than the previous one. Using these notions, we then investigate several fundamental questions, including the existence of EF solutions in the policy teaching setting, the computation of cost-minimizing solutions, and the price of fairness (PoF), i.e., the increase in cost due to consideration of fairness. We show that an EF solution may not exist when penalties are not allowed, but exists otherwise. Depending on the cost measures, computing a cost-minimizing EF solution can be formulated as convex or linear programming and hence solved efficiently. Asymptotically, the PoF increases but at most linearly with the geometric sum of the discount factor in general, the size of the MDP, and the number of agents involved. Thus, fairness can be incorporated in multi-agent teaching without significant computational or price-of-fairness burdens.

## [Continuously Tempered PDMP samplers](#)

- Matthew Sutton Å· Robert Salomone Å· Augustin Chevallier Å· Paul Fearnhead
- abstract@[open-review](#): New sampling algorithms based on simulating continuous-time stochastic processes called piece-wise deterministic Markov processes (PDMPs) have shown considerable promise. However, these methods can struggle to sample from multi-modal or heavy-tailed distributions. We show how tempering ideas can improve the mixing of PDMPs in such cases. We introduce an extended distribution defined over the state of the posterior distribution and an inverse temperature, which interpolates between a tractable distribution when the inverse temperature is 0 and the posterior when the inverse temperature is 1. The marginal distribution of the inverse temperature is a mixture of a continuous distribution on  $[0,1]$  and a point mass at 1: which means that we obtain samples when the inverse temperature is 1, and these are draws from the posterior, but sampling algorithms will also explore distributions at lower temperatures which will improve mixing. We show how PDMPs, and particularly the Zig-Zag sampler, can be implemented to sample from such an extended distribution. The resulting algorithm is easy to implement and we show empirically that it can outperform existing PDMP-based samplers on challenging multimodal posteriors.

## [An Adaptive Kernel Approach to Federated Learning of Heterogeneous Causal Effects](#)

- Thanh Vinh Vo Å· Arnab Bhattacharyya Å· Young Lee Å· Tze-Yun Leong

- abstract@[open-review](#): We propose a new causal inference framework to learn causal effects from multiple, decentralized data sources in a federated setting. We introduce an adaptive transfer algorithm that learns the similarities among the data sources by utilizing Random Fourier Features to disentangle the loss function into multiple components, each of which is associated with a data source. The data sources may have different distributions; the causal effects are independently and systematically incorporated. The proposed method estimates the similarities among the sources through transfer coefficients, and hence requiring no prior information about the similarity measures. The heterogeneous causal effects can be estimated with no sharing of the raw training data among the sources, thus minimizing the risk of privacy leak. We also provide minimax lower bounds to assess the quality of the parameters learned from the disparate sources. The proposed method is empirically shown to outperform the baselines on decentralized data sources with dissimilar distributions.

## [DevFly: Bio-Inspired Development of Binary Connections for Locality Preserving Sparse Codes](#)

- Tianqi Wei Â· Rana Alkhouri Maroun Â· Qinghai Guo Â· Barbara Webb
- abstract@[open-review](#): Neural circuits undergo developmental processes which can be influenced by experience. Here we explore a bio-inspired development process to form the connections in a network used for locality sensitive hashing. The network is a simplified model of the insect mushroom body, which has sparse connections from the input layer to a second layer of higher dimension, forming a sparse code. In previous versions of this model, connectivity between the layers is random. We investigate whether the performance of the hash, evaluated in nearest neighbour query tasks, can be improved by process of developing the connections, in which the strongest input dimensions in successive samples are wired to each successive coding dimension. Experiments show that, the accuracy of searching for nearest neighbours is improved, although performance is dependent on the parameter values and datasets used. Our approach is also much faster than alternative methods that have been proposed for training the connections in this model. Importantly, the development process does not impact connections built at an earlier stage, which should provide stable coding results for simultaneous learning in a downstream network

## [PhysGNN: A Physics--Driven Graph Neural Network Based Model for Predicting Soft Tissue Deformation in Image--Guided Neurosurgery](#)

- Yasmin Salehi Â· Dennis Giannacopoulos
- abstract@[open-review](#): Correctly capturing intraoperative brain shift in image-guided neurosurgical procedures is a critical task for aligning preoperative data with intraoperative geometry for ensuring accurate surgical navigation. While the finite element method (FEM) is a proven technique to effectively approximate soft tissue deformation through biomechanical formulations, their degree of success boils down to a trade-off between accuracy and speed. To circumvent this problem, the most recent works in this domain have proposed leveraging data-driven models obtained by training various machine learning algorithms---e.g. random forests, artificial neural networks (ANNs)---with the results of finite element analysis (FEA) to speed up tissue deformation approximations by prediction. These methods, however, do not account for the structure of the finite element (FE) mesh during training that provides information on node connectivities as well as the distance between them, which can aid with approximating tissue deformation based on the proximity of force load points with the rest of the mesh nodes. Therefore, this work proposes a novel framework, PhysGNN, a data-driven model that approximates the solution of FEA by leveraging graph neural networks (GNNs), which are capable of accounting for the mesh structural information and inductive learning over unstructured grids and complex topological structures. Empirically, we demonstrate that the proposed architecture, PhysGNN, promises accurate and fast soft tissue deformation approximations, and is competitive with the state of the art (SOTA) algorithms while promising enhanced computational feasibility and therefore suitable for neurosurgical settings.

## [Addressing Leakage in Concept Bottleneck Models](#)

- Marton Havasi Â· Sonali Parbhoo Â· Finale Doshi-Velez
- abstract@[open-review](#): Concept bottleneck models (CBMs) enhance the interpretability of their predictions by first predicting high-level concepts given features, and subsequently predicting outcomes on the basis of these concepts. Recently, it was demonstrated that training the label predictor directly on the probabilities produced by the concept predictor as opposed to the ground-truth concepts, improves label predictions. However, this results in corruptions in the concept predictions that impact the concept accuracy as well as our ability to intervene on the concepts -- a key proposed benefit of CBMs. In this work, we investigate and address two issues with CBMs that cause this disparity in performance: having an insufficient concept set and using inexpressive concept predictor. With our modifications, CBMs become competitive in terms of predictive performance, with models that otherwise leak additional information in the concept probabilities, while having dramatically increased concept accuracy and intervention accuracy.

## [Better Uncertainty Calibration via Proper Scores for Classification and Beyond](#)

- Sebastian Gruber Â· Florian Buettner
- abstract@[open-review](#): With model trustworthiness being crucial for sensitive real-world applications, practitioners are putting more and more focus on improving the uncertainty calibration of deep neural networks. Calibration errors are designed to quantify the reliability of probabilistic predictions but their estimators are usually biased and inconsistent. In this work, we introduce the framework of \textit{proper calibration errors}, which relates every calibration error to a proper score and provides a respective upper bound with optimal estimation properties. This relationship can be used to reliably quantify the model calibration improvement. We theoretically and empirically demonstrate the shortcomings of commonly used estimators compared to our approach. Due to the wide applicability of proper scores, this gives a natural extension of recalibration beyond classification.

## [MLA: MultiLingual Acquisition on Multimodal Pre-training](#)

- Liang Zhang Â· Anwen Hu Â· Qin Jin
- abstract@[open-review](#): Vision and diverse languages are important information sources in our living world. A model that understands multi-modalities and multi-languages can be applied to a wider range of real-life scenarios. To build such a multimodal and multilingual model, existing works try to ensemble vision-language data from multiple languages in pre-training. However, due to the large number of languages, these works often require huge computing resources and cannot be flexibly extended to new languages. In this work, we propose a MultiLingual Acquisition (MLA) framework that can easily empower a monolingual Vision-Language Pre-training (VLP) model with multilingual capability. Specifically, we design a lightweight language acquisition encoder based on state-of-the-art monolingual VLP models. We further propose a two-stage training strategy to optimize the language acquisition encoder, namely the Native Language Transfer stage and the Language Exposure stage. With much less multilingual training data and computing resources, our model achieves state-of-the-art performance on multilingual image-text and video-text retrieval benchmarks.

## [Learning on Arbitrary Graph Topologies via Predictive Coding](#)

- Tommaso Salvatori Â· Luca Pinchetti Â· Beren Millidge Â· Yuhang Song Â· Tianyi Bao Â· Rafal Bogacz Â· Thomas Lukasiewicz
- abstract@[open-review](#): Training with backpropagation (BP) in standard deep learning consists of two main steps: a forward pass that maps a data point to its prediction, and a backward pass that propagates the error of this prediction back through the network. This process is highly effective when the goal is to minimize a specific objective function. However, it does not allow training on networks with cyclic or backward connections. This is an obstacle to reaching brain-like capabilities, as the highly complex hierarchical structure of the neural connections in the neocortex are potentially fundamental for its effectiveness. In this paper, we show how predictive coding (PC), a theory of information processing in the cortex, can be used to perform inference and learning on arbitrary graph topologies. We experimentally show how this formulation, called PC graphs, can be used to flexibly perform different tasks with the same network by simply stimulating specific neurons. This enables the model to be queried on stimuli with different structures, such as partial

images, images with labels, or images without labels. We conclude by investigating how the topology of the graph influences the final performance, and comparing against simple baselines trained with BP.

## [Training Scale-Invariant Neural Networks on the Sphere Can Happen in Three Regimes](#)

- Maxim Kodryan · Ekaterina Lobacheva · Maksim Nakhodnov · Dmitry Vetrov
- abstract@[open-review](#): A fundamental property of deep learning normalization techniques, such as batch normalization, is making the pre-normalization parameters scale invariant. The intrinsic domain of such parameters is the unit sphere, and therefore their gradient optimization dynamics can be represented via spherical optimization with varying effective learning rate (ELR), which was studied previously. In this work, we investigate the properties of training scale-invariant neural networks directly on the sphere using a fixed ELR. We discover three regimes of such training depending on the ELR value: convergence, chaotic equilibrium, and divergence. We study these regimes in detail both on a theoretical examination of a toy example and on a thorough empirical analysis of real scale-invariant deep learning models. Each of the regimes possesses its own unique features and has strong parallels with previous research on both general and specific scale-invariant neural networks training. Finally, we demonstrate how the discovered regimes are reflected in the conventional training of normalized networks.

## [A gradient sampling method with complexity guarantees for Lipschitz functions in high and low dimensions](#)

- Damek Davis · Dmitriy Drusvyatskiy · Yin Tat Lee · Swati Padmanabhan · Guanghao Ye
- abstract@[open-review](#): Zhang et al. (ICML 2020) introduced a novel modification of Goldstein's classical subgradient method, with an efficiency guarantee of  $\$O(\sqrt{\epsilon})$  for minimizing Lipschitz functions. Their work, however, makes use of an oracle that is not efficiently implementable. In this paper, we obtain the same efficiency guarantee with a standard subgradient oracle, thus making our algorithm efficiently implementable. Our resulting method works on any Lipschitz function whose value and gradient can be evaluated at points of differentiability. We additionally present a new cutting plane algorithm that achieves an efficiency of  $\$O(d\sqrt{\epsilon}\log S)$  for the class of  $\$S\$$ -smooth (and possibly non-convex) functions in low dimensions. Strikingly, this  $\$S\$$ -dependence matches the lower bounds for the convex setting.

## [On the Representation Collapse of Sparse Mixture of Experts](#)

- Zewen Chi · Li Dong · Shaohan Huang · Damai Dai · Shuming Ma · Barun Patra · Saksham Singhal · Payal Bajaj · XIA SONG · Xian-Ling Mao · Heyan Huang · Furu Wei
- abstract@[open-review](#): Sparse mixture of experts provides larger model capacity while requiring a constant computational overhead. It employs the routing mechanism to distribute input tokens to the best-matched experts according to their hidden representations. However, learning such a routing mechanism encourages token clustering around expert centroids, implying a trend toward representation collapse. In this work, we propose to estimate the routing scores between tokens and experts on a low-dimensional hypersphere. We conduct extensive experiments on cross-lingual language model pre-training and fine-tuning on downstream tasks. Experimental results across seven multilingual benchmarks show that our method achieves consistent gains. We also present a comprehensive analysis on the representation and routing behaviors of our models. Our method alleviates the representation collapse issue and achieves more consistent routing than the baseline mixture-of-experts methods.

## [Certifying Robust Graph Classification under Orthogonal Gromov-Wasserstein Threats](#)

- Hongwei Jin · Zishun Yu · Xinhua Zhang
- abstract@[open-review](#): Graph classifiers are vulnerable to topological attacks. Although certificates of robustness have been recently developed, their threat model only counts local and global edge perturbations, which effectively ignores important graph structures such as isomorphism. To address this issue, we propose measuring the perturbation with the orthogonal Gromov-Wasserstein discrepancy, and building its Fenchel biconjugate to facilitate convex optimization. Our key insight is drawn from the matching loss whose root connects two variables via a monotone operator, and it yields a tight outer convex approximation for resistance distance on graph nodes. When applied to graph classification by graph convolutional networks, both our certificate and attack algorithm are demonstrated effective.

## [Randomized Sketches for Clustering: Fast and Optimal Kernel \$k\$ -Means](#)

- Rong Yin · Yong Liu · Weiping Wang · Dan Meng
- abstract@[open-review](#): Kernel  $k$ -means is arguably one of the most common approaches to clustering. In this paper, we investigate the efficiency of kernel  $k$ -means combined with randomized sketches in terms of both statistical analysis and computational requirements. More precisely, we propose a unified randomized sketches framework to kernel  $k$ -means and investigate its excess risk bounds, obtaining the state-of-the-art risk bound with only a fraction of computations. Indeed, we prove that it suffices to choose the sketch dimension  $\$O(\sqrt{n})$  to obtain the same accuracy of exact kernel  $k$ -means with greatly reducing the computational costs, for sub-Gaussian sketches, the randomized orthogonal system (ROS) sketches, and Nyström kernel  $k$ -means, where  $n$  is the number of samples. To the best of our knowledge, this is the first result of this kind for unsupervised learning. Finally, the numerical experiments on simulated data and real-world datasets validate our theoretical analysis.

## [Learning Rigid Body Dynamics with Lagrangian Graph Neural Network](#)

- Ravinder Bhattoo · Sayan Ranu · N M Anoop Krishnan
- abstract@[open-review](#): Lagrangian and Hamiltonian neural networks (LNN and HNN respectively) encode strong inductive biases that allow them to outperform other models of physical systems significantly. However, these models have, thus far, mostly been limited to simple systems such as pendulums and springs or a single rigid body such as a gyroscope or a rigid rotor. Here, we present a Lagrangian graph neural network (LGNN) that can learn the dynamics of rigid bodies by exploiting their topology. We demonstrate the performance of LGNN by learning the dynamics of ropes, chains, and trusses with the bars modeled as rigid bodies. LGNN also exhibits generalizability—LGNN trained on chains with a few segments exhibits generalizability to simulate a chain with large number of links and arbitrary link length. We also show that the LGNN can simulate unseen hybrid systems including bars and chains, on which they have not been trained on. Specifically, we show that the LGNN can be used to model the dynamics of complex real-world structures such as the stability of tensegrity structures. Finally, we discuss the non-diagonal nature of the mass matrix and its ability to generalize in complex systems.

## [Relaxing Equivariance Constraints with Non-stationary Continuous Filters](#)

- Tycho van der Ouderaa · David W. Romero · Mark van der Wilk
- abstract@[open-review](#): Equivariances provide useful inductive biases in neural network modeling, with the translation equivariance of convolutional neural networks being a canonical example. Equivariances can be embedded in architectures through weight-sharing and place symmetry constraints on the functions a neural network can represent. The type of symmetry is typically fixed and has to be chosen in advance. Although some tasks are inherently equivariant, many tasks do not strictly follow such symmetries. In such cases, equivariance constraints can be overly restrictive. In this work, we propose a parameter-efficient relaxation of equivariance that can effectively interpolate between a (i) non-equivariant linear product, (ii) a strict-equivariant convolution, and (iii) a strictly-invariant mapping. The proposed parameterization can be thought of as a building block to allow adjustable symmetry

structure in neural networks. Compared to non-equivariant or strict-equivariant baselines, we experimentally verify that soft equivariance leads to improved performance in terms of test accuracy on CIFAR-10 and CIFAR-100 image classification tasks.

## [Subquadratic Kronecker Regression with Applications to Tensor Decomposition](#)

- Mehrdad Ghadiri · Matthew Fahrbach · Gang Fu
- abstract@[open-review](#): Kronecker regression is a highly-structured least squares problem  $\min_{\mathbf{x}} \|\mathbf{x}\| - \|\mathbf{K}\mathbf{x} - \mathbf{b}\|^2$ , where the design matrix  $\mathbf{K} = \mathbf{A}^{(1)} \cdots \otimes \mathbf{A}^{(N)}$  is a Kronecker product of factor matrices. This regression problem arises in each step of the widely-used alternating least squares (ALS) algorithm for computing the Tucker decomposition of a tensor. We present the first  $\text{subquadratic-time}$  algorithm for solving Kronecker regression to a  $(1+\varepsilon)$ -approximation that avoids the exponential term  $O(\varepsilon^{-N})$  in the running time. Our techniques combine leverage score sampling and iterative methods. By extending our approach to block-design matrices where one block is a Kronecker product, we also achieve subquadratic-time algorithms for (1) Kronecker ridge regression and (2) updating the factor matrix of a Tucker decomposition in ALS, which is not a pure Kronecker regression problem, thereby improving the running time of all steps of Tucker ALS. We demonstrate the speed and accuracy of this Kronecker regression algorithm on synthetic data and real-world image tensors.

## [CascadeXML: End-to-end Multi-Resolution Learning for Extreme Multi-Label Text Classification](#)

- Siddhant Kharbanda · Atmadeep Banerjee · Erik Schultheis · Rohit Babbar
- abstract@[open-review](#): Extreme Multi-label Text Classification (XMC) involves learning a classifier that can assign an input with a subset of most relevant labels from millions of label choices. Recent approaches, such as XR-Transformer and LightXML, leverage a transformer instance to achieve state-of-the-art performance. However, in this process, these approaches need to make various trade-offs between performance and computational requirements. A major shortcoming, as compared to the Bi-LSTM based AttentionXML, is that they fail to keep separate feature representations for each resolution in a label tree. We thus propose CascadeXML, an end-to-end multi-resolution learning pipeline, which can harness the multi-layered architecture of a transformer model for attending to different label resolutions with separate feature representations. CascadeXML significantly outperforms all existing approaches with non-trivial gains obtained on benchmark datasets consisting of up to three million labels. Code for CascadeXML will be made publicly available.

## [Improved techniques for deterministic \$\ell\_2\$ robustness](#)

- Sahil Singla · Soheil Feizi
- abstract@[open-review](#): Training convolutional neural networks (CNNs) with a strict 1-Lipschitz constraint under the  $\ell_2$  norm is useful for adversarial robustness, interpretable gradients and stable training. 1-Lipschitz CNNs are usually designed by enforcing each layer to have an orthogonal Jacobian matrix (for all inputs) to prevent the gradients from vanishing during backpropagation. However, their performance often significantly lags behind that of heuristic methods to enforce Lipschitz constraints where the resulting CNN is not provably 1-Lipschitz. In this work, we reduce this gap by introducing (a) a procedure to certify robustness of 1-Lipschitz CNNs by replacing the last linear layer with a 1-hidden layer MLP that significantly improves their performance for both standard and provably robust accuracy, (b) a method to significantly reduce the training time per epoch for Skew Orthogonal Convolution (SOC) layers ( $>30\%$  reduction for deeper networks) and (c) a class of pooling layers using the mathematical property that the  $\ell_2$  distance of an input to a manifold is 1-Lipschitz. Using these methods, we significantly advance the state-of-the-art for standard and provable robust accuracies on CIFAR-10 (gains of  $+1.79\%$  and  $+3.82\%$ ) and similarly on CIFAR-100 ( $+3.78\%$  and  $+4.75\%$  across all networks).

## [Unsupervised Adaptation from Repeated Traversals for Autonomous Driving](#)

- Yurong You · Cheng Perng Phoo · Katie Luo · Travis Zhang · Wei-Lun Chao · Bharath Hariharan · Mark Campbell · Kilian Weinberger
- abstract@[open-review](#): For a self-driving car to operate reliably, its perceptual system must generalize to the end-user's environment --- ideally without additional annotation efforts. One potential solution is to leverage unlabeled data (e.g., unlabeled LiDAR point clouds) collected from the end-users' environments (i.e. target domain) to adapt the system to the difference between training and testing environments. While extensive research has been done on such an unsupervised domain adaptation problem, one fundamental problem lingers: there is no reliable signal in the target domain to supervise the adaptation process. To overcome this issue we observe that it is easy to collect unsupervised data from multiple traversals of repeated routes. While different from conventional unsupervised domain adaptation, this assumption is extremely realistic since many drivers share the same roads. We show that this simple additional assumption is sufficient to obtain a potent signal that allows us to perform iterative self-training of 3D object detectors on the target domain. Concretely, we generate pseudo-labels with the out-of-domain detector but reduce false positives by removing detections of supposedly mobile objects that are persistent across traversals. Further, we reduce false negatives by encouraging predictions in regions that are not persistent. We experiment with our approach on two large-scale driving datasets and show remarkable improvement in 3D object detection of cars, pedestrians, and cyclists, bringing us a step closer to generalizable autonomous driving.

## [Byzantine Spectral Ranking](#)

- Arnav Datar · Arun Rajkumar · John Augustine
- abstract@[open-review](#): We study the problem of rank aggregation where the goal is to obtain a global ranking by aggregating pair-wise comparisons of voters over a set of items. We consider an adversarial setting where the voters are partitioned into two sets. The first set votes in a stochastic manner according to the popular score-based Bradley-Terry-Luce (BTL) model for pairwise comparisons. The second set comprises malicious Byzantine voters trying to deteriorate the ranking. We consider a strongly-adversarial scenario where the Byzantine voters know the BTL scores, the votes of the good voters, the algorithm, and can collude with each other. We first show that the popular spectral ranking based Rank-Centrality algorithm, though optimal for the BTL model, does not perform well even when a small constant fraction of the voters are Byzantine. We introduce the Byzantine Spectral Ranking Algorithm (and a faster variant of it), which produces a reliable ranking when the number of good voters exceeds the number of Byzantine voters. We show that no algorithm can produce a satisfactory ranking with probability  $> 1/2$  for all BTL weights when there are more Byzantine voters than good voters, showing that our algorithm works for all possible population fractions. We support our theoretical results with experimental results on synthetic and real datasets to demonstrate the failure of the Rank-Centrality algorithm under several adversarial scenarios and how the proposed Byzantine Spectral Ranking algorithm is robust in obtaining good rankings.

## [On Measuring Excess Capacity in Neural Networks](#)

- Florian Graf · Sebastian Zeng · Bastian Rieck · Marc Niethammer · Roland Kwitt
- abstract@[open-review](#): We study the excess capacity of deep networks in the context of supervised classification. That is, given a capacity measure of the underlying hypothesis class - in our case, empirical Rademacher complexity - by how much can we (a priori) constrain this class while retaining an empirical error on a par with the unconstrained regime? To assess excess capacity in modern architectures (such as residual networks), we extend and unify prior Rademacher complexity bounds to accommodate function composition and addition, as well as the structure of convolutions. The capacity-driving terms in our bounds are the Lipschitz constants of the layers and a  $(2,1)$  group norm distance to the initializations of the convolution weights. Experiments on benchmark datasets of varying task difficulty indicate that (1) there is a substantial amount of excess capacity per task, and (2) capacity can be kept at a surprisingly similar level across tasks. Overall, this suggests a notion of compressibility with respect to weight norms, orthogonal to classic compression via weight pruning.

## [GPT3.int8\(\): 8-bit Matrix Multiplication for Transformers at Scale](#)

- Tim Dettmers · Mike Lewis · Luke Zettlemoyer
- abstract@[open-review](#): Large language models have been widely adopted but require significant GPU memory for inference and finetuning. We develop methods for Int8 matrix multiplication for transformer multi-layer perceptron (MLP) and attention projection layers, which cut the required memory for inference by half while retaining full precision performance. With our method, a 16/32-bit checkpoint can be loaded, converted to Int8, and used immediately without performance degradation -- no post-quantization training is required. The key challenge, which we empirically show for the first time, is that existing quantization methods perform poorly at scale due to emergent outlier feature dimensions. We find that standard quantization techniques for matrix multiplication fail beyond 1.3B parameters. To overcome this barrier, we develop vector-wise quantization, which keeps separate normalization constants for each inner product in the matrix multiplication. Additionally, we identify layer and input invariant feature dimensions in the hidden states, which heavily influence attention and disrupt quantization methods starting at 13B parameters. To scale to 13B, we develop a new mixed-precision matrix decomposition scheme, which allows scaling without performance degradation to at least 13B parameters. This result makes large transformers more accessible, for example, by enabling inference with GPT-J and T5-11B on a single free cloud GPU, GPT-NeoX-20B on a single gaming-grade GPU, and OPT-30B on a single data-center-grade GPU. We open source our software.

## [Bounding and Approximating Intersectional Fairness through Marginal Fairness](#)

- Mathieu Molina · Patrick Loiseau
- abstract@[open-review](#): Discrimination in machine learning often arises along multiple dimensions (a.k.a. protected attributes); it is then desirable to ensure \emph{intersectional fairness}---i.e., that no subgroup is discriminated against. It is known that ensuring \emph{marginal fairness} for every dimension independently is not sufficient in general. Due to the exponential number of subgroups, however, directly measuring intersectional fairness from data is impossible. In this paper, our primary goal is to understand in detail the relationship between marginal and intersectional fairness through statistical analysis. We first identify a set of sufficient conditions under which an exact relationship can be obtained. Then, we prove bounds (easily computable through marginal fairness and other meaningful statistical quantities) in high-probability on intersectional fairness in the general case. Beyond their descriptive value, we show that these theoretical bounds can be leveraged to derive a heuristic improving the approximation and bounds of intersectional fairness by choosing, in a relevant manner, protected attributes for which we describe intersectional subgroups. Finally, we test the performance of our approximations and bounds on real and synthetic data-sets.

## [FR: Folded Rationalization with a Unified Encoder](#)

- Wei Liu · Haozhao Wang · Jun Wang · Ruixuan Li · Chao Yue · YuanKai Zhang
- abstract@[open-review](#): Rationalization aims to strengthen the interpretability of NLP models by extracting a subset of human-intelligible pieces of their inputting texts. Conventional works generally employ a two-phase model in which a generator selects the most important pieces, followed by a predictor that makes predictions based on the selected pieces. However, such a two-phase model may incur the degeneration problem where the predictor overfits to the noise generated by a not yet well-trained generator and in turn, leads the generator to converge to a suboptimal model that tends to select senseless pieces. To tackle this challenge, we propose Folded Rationalization (FR) that folds the two phases of the rationale model into one from the perspective of text semantic extraction. The key idea of FR is to employ a unified encoder between the generator and predictor, based on which FR can facilitate a better predictor by access to valuable information blocked by the generator in the traditional two-phase model and thus bring a better generator. Empirically, we show that FR improves the F1 score by up to 10.3% as compared to state-of-the-art methods.

## [Okapi: Generalising Better by Making Statistical Matches Match](#)

- Myles Bartlett · Sara Romiti · Viktoriia Sharmanska · Novi Quadrianto
- abstract@[open-review](#): We propose Okapi, a simple, efficient, and general method for robust semi-supervised learning based on online statistical matching. Our method uses a nearest-neighbours-based matching procedure to generate cross-domain views for a consistency loss, while eliminating statistical outliers. In order to perform the online matching in a runtime- and memory-efficient way, we draw upon the self-supervised literature and combine a memory bank with a slow-moving momentum encoder. The consistency loss is applied within the feature space, rather than on the predictive distribution, making the method agnostic to both the modality and the task in question. We experiment on the WILDS 2.0 datasets (Sagawa et al.), which significantly expands the range of modalities, applications, and shifts available for studying and benchmarking real-world unsupervised adaptation. Contrary to Sagawa et al., we show that it is in fact possible to leverage additional unlabelled data to improve upon empirical risk minimisation (ERM) results with the right method. Our method outperforms the baseline methods in terms of out-of-distribution (OOD) generalisation on both the iWildCam (a multi-class classification task) and PovertyMap (a regression task) datasets. Furthermore, from a qualitative perspective, we show that the matches produced from the learned encoder are related in semantically meaningful ways.

## [A Multilabel Classification Framework for Approximate Nearest Neighbor Search](#)

- Ville Hyvönen · Elias Jääskeläinen · Teemu Roos
- abstract@[open-review](#): Both supervised and unsupervised machine learning algorithms have been used to learn partition-based index structures for approximate nearest neighbor (ANN) search. Existing supervised algorithms formulate the learning task as finding a partition in which the nearest neighbors of a training set point belong to the same partition element as the point itself, so that the nearest neighbor candidates can be retrieved by naive lookup or backtracking search. We formulate candidate set selection in ANN search directly as a multilabel classification problem where the labels correspond to the nearest neighbors of the query point, and interpret the partitions as partitioning classifiers for solving this task. Empirical results suggest that the natural classifier based on this interpretation leads to strictly improved performance when combined with any unsupervised or supervised partitioning strategy. We also prove a sufficient condition for consistency of a partitioning classifier for ANN search, and illustrate the result by verifying this condition for chronological \$k\$-d trees.

## [\(De-\)Randomized Smoothing for Decision Stump Ensembles](#)

- Miklós Horváth · Mark Müller · Marc Fischer · Martin Vechev
- abstract@[open-review](#): Tree-based models are used in many high-stakes application domains such as finance and medicine, where robustness and interpretability are of utmost importance. Yet, methods for improving and certifying their robustness are severely under-explored, in contrast to those focusing on neural networks. Targeting this important challenge, we propose deterministic smoothing for decision stump ensembles. Whereas most prior work on randomized smoothing focuses on evaluating arbitrary base models approximately under input randomization, the key insight of our work is that decision stump ensembles enable exact yet efficient evaluation via dynamic programming. Importantly, we obtain deterministic robustness certificates, even jointly over numerical and categorical features, a setting ubiquitous in the real world. Further, we derive an MLE-optimal training method for smoothed decision stumps under randomization and propose two boosting approaches to improve their provable robustness. An extensive experimental evaluation shows that our approach yields significantly higher certified accuracies than the state-of-the-art for tree-based models. We release all code and trained models at ANONYMIZED.

## [Power and limitations of single-qubit native quantum neural networks](#)

- Zhan Yu · Hongshun Yao · Mujin Li · Xin Wang
- abstract@[open-review](#): Quantum neural networks (QNNs) have emerged as a leading strategy to establish applications in machine learning, chemistry, and optimization. While the applications of QNN have been widely investigated, its theoretical foundation remains less understood. In this paper, we formulate a theoretical framework for the expressive ability of data re-uploading quantum neural networks that consist of interleaved encoding circuit blocks and trainable circuit blocks. First, we prove that single-qubit quantum neural networks can approximate any univariate function by mapping the model to a partial Fourier series. We in particular establish the exact correlations between the parameters of the trainable gates and the Fourier coefficients, resolving an open problem on the universal approximation property of QNN. Second, we discuss the limitations of single-qubit native QNNs on approximating multivariate functions by analyzing the frequency spectrum and the flexibility of Fourier coefficients. We further demonstrate the expressivity and limitations of single-qubit native QNNs via numerical experiments. We believe these results would improve our understanding of QNNs and provide a helpful guideline for designing powerful QNNs for machine learning tasks.

## [Local Identifiability of Deep ReLU Neural Networks: the Theory](#)

- Joachim Bona-Pellissier · FranÃ§ois Malgouyres · Francois Bachoc
- abstract@[open-review](#): Is a sample rich enough to determine, at least locally, the parameters of a neural network? To answer this question, we introduce a new local parameterization of a given deep ReLU neural network by fixing the values of some of its weights. This allows us to define local lifting operators whose inverses are charts of a smooth manifold of a high dimensional space. The function implemented by the deep ReLU neural network composes the local lifting with a linear operator which depends on the sample. We derive from this convenient representation a geometrical necessary and sufficient condition of local identifiability. Looking at tangent spaces, the geometrical condition provides: 1/ a sharp and testable necessary condition of identifiability and 2/ a sharp and testable sufficient condition of local identifiability. The validity of the conditions can be tested numerically using backpropagation and matrix rank computations.

## [Proximal Point Imitation Learning](#)

- Luca Viano · Angeliki Kamoutsi · Gergely Neu · Igor Krawczuk · Volkan Cevher
- abstract@[open-review](#): This work develops new algorithms with rigorous efficiency guarantees for infinite horizon imitation learning (IL) with linear function approximation without restrictive coherence assumptions. We begin with the minimax formulation of the problem and then outline how to leverage classical tools from optimization, in particular, the proximal-point method (PPM) and dual smoothing, for online and offline IL, respectively. Thanks to PPM, we avoid nested policy evaluation and cost updates for online IL appearing in the prior literature. In particular, we do away with the conventional alternating updates by the optimization of a single convex and smooth objective over both cost and  $\$Q\$$ -functions. When solved inexactly, we relate the optimization errors to the suboptimality of the recovered policy. As an added bonus, by re-interpreting PPM as dual smoothing with the expert policy as a center point, we also obtain an offline IL algorithm enjoying theoretical guarantees in terms of required expert trajectories. Finally, we achieve convincing empirical performance for both linear and neural network function approximation.

## [Joint Entropy Search For Maximally-Informed Bayesian Optimization](#)

- Carl Hvarfner · Frank Hutter · Luigi Nardi
- abstract@[open-review](#): Information-theoretic Bayesian optimization techniques have become popular for optimizing expensive-to-evaluate black-box functions due to their non-myopic qualities. Entropy Search and Predictive Entropy Search both consider the entropy over the optimum in the input space, while the recent Max-value Entropy Search considers the entropy over the optimal value in the output space. We propose Joint Entropy Search (JES), a novel information-theoretic acquisition function that considers an entirely new quantity, namely the entropy over the joint optimal probability density over both input and output space. To incorporate this information, we consider the reduction in entropy from conditioning on fantasized optimal input/output pairs. The resulting approach primarily relies on standard GP machinery and removes complex approximations typically associated with information-theoretic methods. With minimal computational overhead, JES shows superior decision-making, and yields state-of-the-art performance for information-theoretic approaches across a wide suite of tasks. As a light-weight approach with superior results, JES provides a new go-to acquisition function for Bayesian optimization.

## [Efficient Meta Reinforcement Learning for Preference-based Fast Adaptation](#)

- Zhizhou Ren · Anji Liu · Yitao Liang · Jian Peng · Jianzhu Ma
- abstract@[open-review](#): Learning new task-specific skills from a few trials is a fundamental challenge for artificial intelligence. Meta reinforcement learning (meta-RL) tackles this problem by learning transferable policies that support few-shot adaptation to unseen tasks. Despite recent advances in meta-RL, most existing methods require the access to the environmental reward function of new tasks to infer the task objective, which is not realistic in many practical applications. To bridge this gap, we study the problem of few-shot adaptation in the context of human-in-the-loop reinforcement learning. We develop a meta-RL algorithm that enables fast policy adaptation with preference-based feedback. The agent can adapt to new tasks by querying human's preference between behavior trajectories instead of using per-step numeric rewards. By extending techniques from information theory, our approach can design query sequences to maximize the information gain from human interactions while tolerating the inherent error of non-expert human oracle. In experiments, we extensively evaluate our method, Adaptation with Noisy Oracle (ANOLOE), on a variety of meta-RL benchmark tasks and demonstrate substantial improvement over baseline algorithms in terms of both feedback efficiency and error tolerance.

## [Active Exploration for Inverse Reinforcement Learning](#)

- David Lindner · Andreas Krause · Giorgia Ramponi
- abstract@[open-review](#): Inverse Reinforcement Learning (IRL) is a powerful paradigm for inferring a reward function from expert demonstrations. Many IRL algorithms require a known transition model and sometimes even a known expert policy, or they at least require access to a generative model. However, these assumptions are too strong for many real-world applications, where the environment can be accessed only through sequential interaction. We propose a novel IRL algorithm: Active exploration for Inverse Reinforcement Learning (AceIRL), which actively explores an unknown environment and expert policy to quickly learn the expert's reward function and identify a good policy. AceIRL uses previous observations to construct confidence intervals that capture plausible reward functions and find exploration policies that focus on the most informative regions of the environment. AceIRL is the first approach to active IRL with sample-complexity bounds that does not require a generative model of the environment. AceIRL matches the sample complexity of active IRL with a generative model in the worst case. Additionally, we establish a problem-dependent bound that relates the sample complexity of AceIRL to the suboptimality gap of a given IRL problem. We empirically evaluate AceIRL in simulations and find that it significantly outperforms more naive exploration strategies.

## [Generalization Analysis of Message Passing Neural Networks on Large Random Graphs](#)

- Sohir Maskey · Ron Levie · Yunseok Lee · Gitta Kutyniok
- abstract@[open-review](#): Message passing neural networks (MPNN) have seen a steep rise in popularity since their introduction as generalizations of convolutional neural networks to graph-structured data, and are now considered state-of-the-art tools for solving a large variety of graph-focused problems. We study the generalization error of MPNNs in graph classification and regression. We assume that graphs of different classes are sampled from different random graph models. We show that, when training a MPNN on a dataset sampled from such a distribution, the generalization gap increases in the complexity of the MPNN, and decreases, not only with respect to the number of training samples, but also with the average number of nodes in the

graphs. This shows how a MPNN with high complexity can generalize from a small dataset of graphs, as long as the graphs are large. The generalization bound is derived from a uniform convergence result, that shows that any MPNN, applied on a graph, approximates the MPNN applied on the geometric model that the graph discretizes.

## [Isometric 3D Adversarial Examples in the Physical World](#)

- yibo miao · Yinpeng Dong · Jun Zhu · Xiao-Shan Gao
- abstract@[open-review](#): Recently, several attempts have demonstrated that 3D deep learning models are as vulnerable to adversarial example attacks as 2D models. However, these methods are still far from stealthy and suffer from severe performance degradation in the physical world. Although 3D data is highly structured, it is difficult to bound the perturbations with simple metrics in the Euclidean space. In this paper, we propose a novel  $\$epsilon$ -isometric ( $\$epsilon$ -ISO) attack method to generate natural and robust 3D adversarial examples in the physical world by considering the geometric properties of 3D objects and the invariance to physical transformations. For naturalness, we constrain the adversarial example and the original one to be  $\$epsilon$ -isometric by adopting the Gaussian curvature as the surrogate metric under a theoretical analysis. For robustness under physical transformations, we propose a maxima over transformation (MaxOT) method to actively search for the most difficult transformations rather than random ones to make the generated adversarial example more robust in the physical world. Extensive experiments on typical point cloud recognition models validate that our approach can improve the attack success rate and naturalness of the generated 3D adversarial examples than the state-of-the-art attack methods.

## [Exploitability Minimization in Games and Beyond](#)

- Denizalp Goktas · Amy Greenwald
- abstract@[open-review](#): Pseudo-games are a natural and well-known generalization of normal-form games, in which the actions taken by each player affect not only the other players' payoffs, as in games, but also the other players' strategy sets. The solution concept par excellence for pseudo-games is the generalized Nash equilibrium (GNE), i.e., a strategy profile at which each player's strategy is feasible and no player can improve their payoffs by unilaterally deviating to another strategy in the strategy set determined by the other players' strategies. The computation of GNE in pseudo-games has long been a problem of interest, due to applications in a wide variety of fields, from environmental protection to logistics to telecommunications. Although the computation of GNE is PPAD-hard in general, it is still of interest to try to compute them in restricted classes of pseudo-games. The literature thus far has focused on asymptotic convergence of search procedures; there are very few, if any, results on the computational complexity of GNE. In this paper, we develop fast exploitability-minimization methods that compute exact or approximate GNE in pseudo-games with jointly convex constraints. We derive convergence guarantees for our methods, and we demonstrate their superiority in experiments over a baseline algorithm for a variety of benchmark pseudo-games.

## [Policy Optimization with Linear Temporal Logic Constraints](#)

- Cameron Voloshin · Hoang Le · Swarat Chaudhuri · Yisong Yue
- abstract@[open-review](#): We study the problem of policy optimization (PO) with linear temporal logic (LTL) constraints. The language of LTL allows flexible description of tasks that may be unnatural to encode as a scalar cost function. We consider LTL-constrained PO as a systematic framework, decoupling task specification from policy selection, and an alternative to the standard of cost shaping. With access to a generative model, we develop a model-based approach that enjoys a sample complexity analysis for guaranteeing both task satisfaction and cost optimality (through a reduction to a reachability problem). Empirically, our algorithm can achieve strong performance even in low sample regimes.

## [Bring Your Own Algorithm for Optimal Differentially Private Stochastic Minimax Optimization](#)

- Liang Zhang · Kiran Thekumparampil · Sewoong Oh · Niao He
- abstract@[open-review](#): We study differentially private (DP) algorithms for smooth stochastic minimax optimization, with stochastic minimization as a byproduct. The holy grail of these settings is to guarantee the optimal trade-off between the privacy and the excess population loss, using an algorithm with a linear time-complexity in the number of training samples. We provide a general framework for solving differentially private stochastic minimax optimization (DP-SMO) problems, such that practitioners can bring their own base optimization algorithm and take it as a black-box to obtain the near-optimal privacy-loss trade-off. Our framework is inspired from the recently proposed Phased-ERM method [21] for nonsmooth differentially private stochastic convex optimization (DP-SCO), which exploits the stability of the empirical risk minimization (ERM) for the privacy guarantee. The flexibility of our approach enables us to sidestep the requirement that the base algorithm needs to have bounded sensitivity, and allows the use of sophisticated variance-reduced accelerated methods to achieve near-linear time-complexity. To the best of our knowledge, these are the first near-linear time algorithms with near-optimal guarantees on the population duality gap for smooth DP-SMO, when the objective is (strongly-)convex“(strongly-)concave. Additionally, based on our flexible framework, we enrich the family of near-linear time algorithms for smooth DP-SCO with near-optimal privacy-loss trade-offs.

## [HyperDomainNet: Universal Domain Adaptation for Generative Adversarial Networks](#)

- Aibek Alanov · Vadim Titov · Dmitry Vetrov
- abstract@[open-review](#): Domain adaptation framework of GANs has achieved great progress in recent years as a main successful approach of training contemporary GANs in the case of very limited training data. In this work, we significantly improve this framework by proposing an extremely compact parameter space for fine-tuning the generator. We introduce a novel domain-modulation technique that allows to optimize only 6 thousand-dimensional vector instead of 30 million weights of StyleGAN2 to adapt to a target domain. We apply this parameterization to the state-of-art domain adaptation methods and show that it has almost the same expressiveness as the full parameter space. Additionally, we propose a new regularization loss that considerably enhances the diversity of the fine-tuned generator. Inspired by the reduction in the size of the optimizing parameter space we consider the problem of multi-domain adaptation of GANs, i.e. setting when the same model can adapt to several domains depending on the input query. We propose the HyperDomainNet that is a hypernetwork that predicts our parameterization given the target domain. We empirically confirm that it can successfully learn a number of domains at once and may even generalize to unseen domains.

## [Off-Policy Evaluation with Deficient Support Using Side Information](#)

- Nicolò Felicioni · Maurizio Ferrari Dacrema · Marcello Restelli · Paolo Cremonesi
- abstract@[open-review](#): The Off-Policy Evaluation (OPE) problem consists in evaluating the performance of new policies from the data collected by another one. OPE is crucial when evaluating a new policy online is too expensive or risky. Many of the state-of-the-art OPE estimators are based on the Inverse Propensity Scoring (IPS) technique, which provides an unbiased estimator when the full support assumption holds, i.e., when the logging policy assigns a non-zero probability to each action. However, there are several scenarios where this assumption does not hold in practice, i.e., there is deficient support, and the IPS estimator is biased in the general case. In this paper, we consider two alternative estimators for the deficient support OPE problem. We first show how to adapt an estimator that was originally proposed for a different domain to the deficient support setting. Then, we propose another estimator, which is a novel contribution of this paper. These estimators exploit additional information about the actions, which we call side information, in order to make reliable estimates on the unsupported actions. Under alternative assumptions that do not require full support, we show that the considered estimators are unbiased. We also provide a theoretical analysis of the concentration when relaxing all the assumptions. Finally, we provide an experimental evaluation showing how the considered estimators are better suited for the deficient support setting than the IPS baseline.

## [Measures of Information Reflect Memorization Patterns](#)

- Rachit Bansal · Danish Pruthi · Yonatan Belinkov
- abstract@[open-review](#): Neural networks are known to exploit spurious artifacts (or shortcuts) that co-occur with a target label, exhibiting heuristic memorization. On the other hand, networks have been shown to memorize training examples, resulting in example-level memorization. These kinds of memorization impede generalization of networks beyond their training distributions. Detecting such memorization could be challenging, often requiring researchers to curate tailored test sets. In this work, we hypothesize and subsequently show that the diversity in the activation patterns of different neurons is reflective of model generalization and memorization. We quantify the diversity in the neural activations through information-theoretic measures and find support for our hypothesis on experiments spanning several natural language and vision tasks. Importantly, we discover that information organization points to the two forms of memorization, even for neural activations computed on unlabeled in-distribution examples. Lastly, we demonstrate the utility of our findings for the problem of model selection.

## [Unlabelled Sample Compression Schemes for Intersection-Closed Classes and Extremal Classes](#)

- Joachim Rubinstein · Benjamin Rubinstein
- abstract@[open-review](#): The sample compressibility of concept classes plays an important role in learning theory, as a sufficient condition for PAC learnability, and more recently as an avenue for robust generalisation in adaptive data analysis. Whether compression schemes of size  $O(d)$  must necessarily exist for all classes of VC dimension  $d$  is unknown, but conjectured to be true by Warmuth. Recently Chalopin, Chepoi, Moran, and Warmuth (2018) gave a beautiful unlabelled sample compression scheme of size VC dimension for all maximum classes: classes that meet the Sauer-Shelah-Perles Lemma with equality. They also offered a counterexample to compression schemes based on a promising approach known as corner peeling. In this paper we simplify and extend their proof technique to deal with so-called extremal classes of VC dimension  $d$  which contain maximum classes of VC dimension  $d-1$ . A criterion is given which would imply that all extremal classes admit unlabelled compression schemes of size  $d$ . We also prove that all intersection-closed classes with VC dimension  $d$  admit unlabelled compression schemes of size at most  $11d$ .

## [Collaborative Learning by Detecting Collaboration Partners](#)

- Shu Ding · Wei Wang
- abstract@[open-review](#): Massive amounts of data are naturally dispersed over different clients in many real-world applications, collaborative learning has been a promising paradigm that allows to learn models through collaboration among the clients. However, leveraging these dispersed data to learn good models is still challenging since data over different clients are heterogeneous. Previous works mainly focus on learning the centralized model for all clients or learning a personalized model for each client. When there are numerous clients, the centralized model performs badly on some clients, while learning a personalized model for each client costs unaffordable computational resources. In this paper, we propose the collaborative learning method to detect collaboration partners and adaptively learn  $K$  models for numerous heterogeneous clients. We theoretically prove that the model learned for each client is a good approximation of its personalized model. Experimental results on real-world datasets verify the effectiveness of our method.

## [Batch Bayesian optimisation via density-ratio estimation with guarantees](#)

- Rafael Oliveira · Louis Tiao · Fabio Ramos
- abstract@[open-review](#): Bayesian optimisation (BO) algorithms have shown remarkable success in applications involving expensive black-box functions. Traditionally BO has been set as a sequential decision-making process which estimates the utility of query points via an acquisition function and a prior over functions, such as a Gaussian process. Recently, however, a reformulation of BO via density-ratio estimation (BORE) allowed reinterpreting the acquisition function as a probabilistic binary classifier, removing the need for an explicit prior over functions and increasing scalability. In this paper, we present a theoretical analysis of BORE's regret and an extension of the algorithm with improved uncertainty estimates. We also show that BORE can be naturally extended to a batch optimisation setting by recasting the problem as approximate Bayesian inference. The resulting algorithm comes equipped with theoretical performance guarantees and is assessed against other batch BO baselines in a series of experiments.

## [Asynchronous SGD Beats Minibatch SGD Under Arbitrary Delays](#)

- Blake Woodworth · Mathieu Even · Konstantin Mishchenko · Francis Bach
- abstract@[open-review](#): The existing analysis of asynchronous stochastic gradient descent (SGD) degrades dramatically when any delay is large, giving the impression that performance depends primarily on the delay. On the contrary, we prove much better guarantees for the same asynchronous SGD algorithm regardless of the delays in the gradients, depending instead just on the number of parallel devices used to implement the algorithm. Our guarantees are strictly better than the existing analyses, and we also argue that asynchronous SGD outperforms synchronous minibatch SGD in the settings we consider. For our analysis, we introduce a novel recursion based on ``virtual iterates'' and delay-adaptive stepsizes, which allow us to derive state-of-the-art guarantees for both convex and non-convex objectives.

## [Active Learning with Neural Networks: Insights from Nonparametric Statistics](#)

- Yinglun Zhu · Robert Nowak
- abstract@[open-review](#): Deep neural networks have great representation power, but typically require large numbers of training examples. This motivates deep active learning methods that can significantly reduce the amount of labeled training data. Empirical successes of deep active learning have been recently reported in the literature, however, rigorous label complexity guarantees of deep active learning have remained elusive. This constitutes a significant gap between theory and practice. This paper tackles this gap by providing the first near minimax optimal label complexity guarantees for deep active learning. The key insight is to study deep active learning from the nonparametric classification perspective. Under common low noise conditions, we show that active learning with neural networks can provably achieve the minimax label complexity, up to disagreement coefficient and other logarithmic terms. When equipped with an abstention option, we further develop a computationally efficient deep active learning algorithm that guarantees  $\mathcal{O}(\frac{1}{\epsilon})$  label complexity, without any low noise assumptions. We also provide extensions of results beyond the commonly studied Sobolev/Hölder spaces and develop label complexity guarantees for learning in Radon  $L^2$  spaces, which have recently been proposed as the natural function spaces associated with neural networks.

## [An In-depth Study of Stochastic Backpropagation](#)

- Jun Fang · Mingze Xu · Hao Chen · Bing Shuai · Zhuowen Tu · Joseph Tighe
- abstract@[open-review](#): In this paper, we provide an in-depth study of Stochastic Backpropagation (SBP) when training deep neural networks for standard image classification and object detection tasks. During backward propagation, SBP calculates the gradients by only using a subset of the feature maps to save memory and the computation cost. We interpret SBP as an efficient way to implement stochastic gradient decent by performing backpropagation dropout, which leads to considerable memory saving and training process speedup, with a minimal impact on the overall model accuracy. We offer a good practice to apply SBP in training deep image classification models, which can be adopted in learning a wide range of deep neural networks. Experiments on image classification and object detection show that SBP can save up to 40% of GPU memory with less than 1% accuracy degradation.

## [A time-resolved theory of information encoding in recurrent neural networks](#)

- Rainer Engelken Å· Sven Goedeke
- abstract@[open-review](#): Mammalian brains process information by the collective dynamics of a deeply layered structure of recurrent networks. Information transmission in neural circuits depends on how well time-varying stimuli are encoded by neural populations. A dynamic balance of externally incoming currents by strong recurrent inhibition was previously proposed as a mechanism to accurately and robustly encode the information in a time-varying stimulus, but a full theory was missing. Here, we develop a non-stationary dynamic mean-field theory that transparently explains how tight balance of excitatory currents by recurrent inhibition improves information encoding. We demonstrate that the mutual information rate of a time-varying input increases linearly with the tightness of balance, both in the presence of additive noise and with recurrent chaotic network fluctuations. We corroborated our findings in numerical experiments and demonstrated that recurrent networks with positive firing rates trained to transmit a time-varying stimulus generically use recurrent inhibition to increase the information rate. We also found that networks trained to transmit multiple independent time-varying signals spontaneously form multiple local inhibitory clusters, one for each input channel. Our findings suggest that feedforward excitatory input and local recurrent inhibition - as observed in many biological circuits - is a generic circuit motif for encoding and transmitting time-varying information in recurrent neural circuits.

## [Private Set Generation with Discriminative Information](#)

- Dingfan Chen Å· Raouf Kerkouche Å· Mario Fritz
- abstract@[open-review](#): Differentially private data generation techniques have become a promising solution to the data privacy challenge â€“ it enables sharing of data while complying with rigorous privacy guarantees, which is essential for scientific progress in sensitive domains. Unfortunately, restricted by the inherent complexity of modeling high-dimensional distributions, existing private generative models are struggling with the utility of synthetic samples. In contrast to existing works that aim at fitting the complete data distribution, we directly optimize for a small set of samples that are representative of the distribution, which is generally an easier task and more suitable for private training. Moreover, we exploit discriminative information from downstream tasks to further ease the training. Our work provides an alternative view for differentially private generation of high-dimensional data and introduces a simple yet effective method that greatly improves the sample utility of state-of-the-art approaches.

## [Equivariant Networks for Zero-Shot Coordination](#)

- Darius Muglich Å· Christian Schroeder de Witt Å· Elise van der Pol Å· Shimon Whiteson Å· Jakob Foerster
- abstract@[open-review](#): Successful coordination in Dec-POMDPs requires agents to adopt robust strategies and interpretable styles of play for their partner. A common failure mode is symmetry breaking, when agents arbitrarily converge on one out of many equivalent but mutually incompatible policies. Commonly these examples include partial observability, e.g. waving your right hand vs. left hand to convey a covert message. In this paper, we present a novel equivariant network architecture for use in Dec-POMDPs that prevents the agent from learning policies which break symmetries, doing so more effectively than prior methods. Our method also acts as a "coordination-improvement operator" for generic, pre-trained policies, and thus may be applied at test-time in conjunction with any self-play algorithm. We provide theoretical guarantees of our work and test on the AI benchmark task of Hanabi, where we demonstrate our methods outperforming other symmetry-aware baselines in zero-shot coordination, as well as able to improve the coordination ability of a variety of pre-trained policies. In particular, we show our method can be used to improve on the state of the art for zero-shot coordination on the Hanabi benchmark.

## [Grounded Reinforcement Learning: Learning to Win the Game under Human Commands](#)

- Shusheng Xu Å· Huaijie Wang Å· YI WU
- abstract@[open-review](#): We consider the problem of building a reinforcement learning (RL) agent that can both accomplish non-trivial tasks, like winning a real-time strategy game, and strictly follow high-level language commands from humans, like â€œattackâ€, even if a command is sub-optimal. We call this novel yet important problem, Grounded Reinforcement Learning (GRL). Compared with other language grounding tasks, GRL is particularly non-trivial and cannot be simply solved by pure RL or behavior cloning (BC). From the RL perspective, it is extremely challenging to derive a precise reward function for human preferences since the commands are abstract and the valid behaviors are highly complicated and multi-modal. From the BC perspective, it is impossible to obtain perfect demonstrations since human strategies in complex games are typically sub-optimal. We tackle GRL via a simple, tractable, and practical constrained RL objective and develop an iterative RL algorithm, REinforced demonstration Distillation (RED), to obtain a strong GRL policy. We evaluate the policies derived by RED, BC and pure RL methods on a simplified real-time strategy game, MiniRTS. Experiment results and human studies show that the RED policy is able to consistently follow human commands and achieve a higher win rate than the baselines.

## [Tractable Optimality in Episodic Latent MABs](#)

- Jeongyeol Kwon Å· Yonathan Efroni Å· Constantine Caramanis Å· Shie Mannor
- abstract@[open-review](#): We consider a multi-armed bandit problem with \$M\$ latent contexts, where an agent interacts with the environment for an episode of \$H\$ time steps. Depending on the length of the episode, the learner may not be able to estimate accurately the latent context. The resulting partial observation of the environment makes the learning task significantly more challenging. Without any additional structural assumptions, existing techniques to tackle partially observed settings imply the decision maker can learn a near-optimal policy with \$O(A)^H\$ episodes, but do not promise more. In this work, we show that learning with \$\{em polynomial\}\$ samples in \$A\$ is possible. We achieve this by using techniques from experiment design. Then, through a method-of-moments approach, we design a procedure that provably learns a near-optimal policy with \$O(\text{poly}(A) + \text{poly}(M,H)^{\lfloor \min(M,H) \rfloor})\$ interactions. In practice, we show that we can formulate the moment-matching via maximum likelihood estimation. In our experiments, this significantly outperforms the worst-case guarantees, as well as existing practical methods.

## [Phase diagram of Stochastic Gradient Descent in high-dimensional two-layer neural networks](#)

- Rodrigo Veiga Å· Ludovic Stephan Å· Bruno Loureiro Å· Florent Krzakala Å· Lenka ZdeborovÃ¡;
- abstract@[open-review](#): Despite the non-convex optimization landscape, over-parametrized shallow networks are able to achieve global convergence under gradient descent. The picture can be radically different for narrow networks, which tend to get stuck in badly-generalizing local minima. Here we investigate the cross-over between these two regimes in the high-dimensional setting, and in particular investigate the connection between the so-called mean-field/hydrodynamic regime and the seminal approach of Saad \& Solla. Focusing on the case of Gaussian data, we study the interplay between the learning rate, the time scale, and the number of hidden units in the high-dimensional dynamics of stochastic gradient descent (SGD). Our work builds on a deterministic description of SGD in high-dimensions from statistical physics, which we extend and for which we provide rigorous convergence rates.

## [Patching open-vocabulary models by interpolating weights](#)

- Gabriel Ilharco Å· Mitchell Wortsman Å· Samir Yitzhak Gadre Å· Shuran Song Å· Hannaneh Hajishirzi Å· Simon Kornblith Å· Ali Farhadi Å· Ludwig Schmidt
- abstract@[open-review](#): While open-vocabulary models like CLIP achieve high accuracy across many image classification tasks, there are still settings where their zero-shot performance is far from optimal. We study model patching, where the goal is to improve accuracy on specific tasks (i.e., patching tasks) without degrading accuracy on tasks where performance is already adequate (i.e., supported tasks). Given a task to be patched, we first fine-tune without introducing new parameters, then interpolate the fine-tuned model weights with the model weights before fine-tuning. We explore model patching on nine tasks where zero-shot CLIP performs poorly, observing that patching increases accuracy by 15 to 60 percentage points while preserving accuracy on ImageNet within one percentage point of the zero-shot model. Additionally, we find that patching is more effective for larger models, demonstrate that

a single model can be patched on multiple tasks, and identify cases of broad transfer, where patching one task can increase accuracy on other tasks even when the classes are not shared. Finally, we investigate applications beyond standard benchmarks, including a patch which makes CLIP less susceptible to typographic attacks. Our findings demonstrate that it is possible to expand the set of tasks on which open-vocabulary models achieve high accuracy without re-training them from scratch.

## [Differentially Private Learning Needs Hidden State \(Or Much Faster Convergence\)](#)

- Jiayuan Ye · Reza Shokri
- abstract@[open-review](#): Differential privacy analysis of randomized learning algorithms typically relies on composition theorems, where the implicit assumption is that the internal state of the iterative algorithm is revealed to the adversary. However, by assuming that the internal state of the algorithm is not revealed, recent works prove a smaller privacy bound for noisy gradient descent (on strongly convex smooth loss functions), compared with composition bounds. In this paper, we significantly improve privacy analysis under the hidden state assumption. We enable taking advantage of privacy amplification by sub-sampling and randomized post-processing, and extend the analysis to shuffle and partition "and sample without replacement" stochastic minibatch gradient descent schemes. We prove that, in these settings, our privacy bound is substantially smaller than the composition bounds, notably after many epochs of training (required for learning from high-dimensional data). So, unless the DP algorithm converges fast, our privacy analysis shows that differentially private learning benefits from hidden state privacy analysis. We also empirically show that, given a fixed privacy budget, our analysis enables DP training of image classification models with a better prediction accuracy, compared with prior work.

## [Additive MIL: Intrinsically Interpretable Models for Digital Pathology](#)

- Syed Ashar Javed · Dinkar Juyal · Harshith Padigela · Amaro Taylor-Weiner · Limin Yu · Aaditya Prakash
- abstract@[open-review](#): Multiple Instance Learning (MIL) has been widely applied in pathology towards solving critical problems such as automating cancer diagnosis and grading, predicting patient prognosis, and therapy response. Deploying these models in a clinical setting requires careful inspection of these black boxes during development and deployment to identify failures and maintain physician trust. In this work, we propose a simple formulation of MIL models, which enables interpretability while maintaining similar predictive performance. Our Additive MIL models enable spatial credit assignment such that the contribution of each region in the image can be exactly computed and visualized. We show that our spatial credit assignment coincides with regions used by pathologists during diagnosis and improves upon classical attention heatmaps from attention MIL models. We show that any existing MIL model can be made additive with a simple change in function composition. We also show how these models can debug model failures, identify spurious features, and highlight class-wise regions of interest, enabling their use in high-stakes environments such as clinical decision-making.

## [On Leave-One-Out Conditional Mutual Information For Generalization](#)

- Mohamad Rida Rammal · Alessandro Achille · Suhas Diggavi · Stefano Soatto · Aditya Golatkar
- abstract@[open-review](#): We derive information theoretic generalization bounds for supervised learning algorithms based on a new measure of leave-one-out conditional mutual information (loo-CMI). Contrary to other CMI bounds, which are black-box bounds that do not exploit the structure of the problem and may be hard to evaluate in practice, our loo-CMI bounds can be computed easily and can be interpreted in connections to other notions such as classical leave-one-out cross-validation, stability of the optimization algorithm and the geometry of the loss-landscape. It applies both to the output of training algorithms as well as their predictions. We empirically validate the quality of the bound by evaluating its predicted generalization gap in scenarios for deep learning. In particular, our bounds are non-vacuous on large-scale image-classification tasks.

## [Truly Deterministic Policy Optimization](#)

- Ehsan Saleh · Saba Ghaffari · Tim Bretl · Matthew West
- abstract@[open-review](#): In this paper, we present a policy gradient method that avoids exploratory noise injection and performs policy search over the deterministic landscape, with the goal of improving learning with long horizons and non-local rewards. By avoiding noise injection all sources of estimation variance can be eliminated in systems with deterministic dynamics (up to the initial state distribution). Since deterministic policy regularization is impossible using traditional non-metric measures such as the KL divergence, we derive a Wasserstein-based quadratic model for our purposes. We state conditions on the system model under which it is possible to establish a monotonic policy improvement guarantee, propose a surrogate function for policy gradient estimation, and show that it is possible to compute exact advantage estimates if both the state transition model and the policy are deterministic. Finally, we describe two novel robotic control environments---one with non-local rewards in the frequency domain and the other with a long horizon (8000 time-steps)---for which our policy gradient method (TDPO) significantly outperforms existing methods (PPO, TRPO, DDPG, and TD3). Our implementation with all the experimental settings is available at [https://anonymous.4open.science/r/code\\_tdpo-D23A](https://anonymous.4open.science/r/code_tdpo-D23A).

## [DiSC: Differential Spectral Clustering of Features](#)

- Ram Dyuthi Sristi · Gal Mishne · Ariel Jaffe
- abstract@[open-review](#): Selecting subsets of features that differentiate between two conditions is a key task in a broad range of scientific domains. In many applications, the features of interest form clusters with similar effects on the data at hand. To recover such clusters we develop DiSC, a data-driven approach for detecting groups of features that differentiate between conditions. For each condition, we construct a graph whose nodes correspond to the features and whose weights are functions of the similarity between them for that condition. We then apply a spectral approach to compute subsets of nodes whose connectivity differs significantly between the condition-specific feature graphs. On the theoretical front, we analyze our approach with a toy example based on the stochastic block model. We evaluate DiSC on a variety of datasets, including MNIST, hyperspectral imaging, simulated scRNA-seq and task fMRI, and demonstrate that DiSC uncovers features that better differentiate between conditions compared to competing methods.

## [You Can't Count on Luck: Why Decision Transformers Fail in Stochastic Environments](#)

- Keiran Paster · Sheila McIlraith · Jimmy Ba
- abstract@[open-review](#): Recently, methods such as Decision Transformer that reduce reinforcement learning to a prediction task and solve it via supervised learning (RvS) have become popular due to their simplicity, robustness to hyperparameters, and strong overall performance on offline RL tasks. However, simply conditioning a probabilistic model on a desired return and taking the predicted action can fail dramatically in stochastic environments since trajectories that result in a return may have only achieved that return due to luck. In this work, we describe the limitations of RvS approaches in stochastic environments and propose a solution. Rather than simply conditioning on returns, as is standard practice, our proposed method, ESPER, conditions on learned average returns which are independent from environment stochasticity. Doing so allows ESPER to achieve strong alignment between target return and expected performance in real environments. We demonstrate this in several challenging stochastic offline-RL tasks including the challenging puzzle game 2048, and Connect Four playing against a stochastic opponent. In all tested domains, ESPER achieves significantly better alignment between the target return and achieved return than simply conditioning on returns. ESPER also achieves higher maximum performance than even the value-based baselines.

## [No Free Lunch from Deep Learning in Neuroscience: A Case Study through Models of the Entorhinal-Hippocampal Circuit](#)

- Rylan Schaeffer · Mikail Khona · Ila Fiete

- abstract@[open-review](#): Research in Neuroscience, as in many scientific disciplines, is undergoing a renaissance based on deep learning. Unique to Neuroscience, deep learning models can be used not only as a tool but interpreted as models of the brain. The central claims of recent deep learning-based models of brain circuits are that they make novel predictions about neural phenomena or shed light on the fundamental functions being optimized. We show, through the case-study of grid cells in the entorhinal-hippocampal circuit, that one may get neither. We begin by reviewing the principles of grid cell mechanism and function obtained from first-principles modeling efforts, then rigorously examine the claims of deep learning models of grid cells. Using large-scale hyperparameter sweeps and theory-driven experimentation, we demonstrate that the results of such models may be more strongly driven by particular, non-fundamental, and post-hoc implementation choices than fundamental truths about neural circuits or the loss function(s) they might optimize. We discuss why these models cannot be expected to produce accurate models of the brain without the addition of substantial amounts of inductive bias, an informal No Free Lunch result for Neuroscience. Based on first principles work, we provide hypotheses for what additional loss functions will produce grid cells more robustly. In conclusion, caution and consideration, together with biological knowledge, are warranted in building and interpreting deep learning models in Neuroscience.

## [Zero-shot Transfer Learning on Heterogeneous Graphs via Knowledge Transfer Networks](#)

- Minji Yoon · John Palowitch · Dustin Zelle · Ziniu Hu · Ruslan Salakhutdinov · Bryan Perozzi
- abstract@[open-review](#): Data continuously emitted from industrial ecosystems such as social or commerce platforms are commonly represented as heterogeneous graphs (HG) composed of multiple node/edge types. State-of-the-art graph learning methods for HG known as heterogeneous graph neural networks (HGN) are applied to learn deep context-informed node representations. However, many HG datasets from industrial applications suffer from label imbalance between node types. As there is no direct way to learn using labels rooted at different node types, HGNs have been applied on only a few node types with abundant labels. We propose a zero-shot transfer learning module for HGNs called a Knowledge Transfer Network (KTN) that transfers knowledge from label-abundant node types to zero-labeled node types through rich relational information given in the HG. KTN is derived from the theoretical relationship between distinct feature extractors for each node type given in the HGNs, which we introduce in this work. KTN improves the performance of 6 different types of HGN models up to 960% for inference on zero-labeled node types and outperforms state-of-the-art transfer learning baselines up to 73% across 18 different transfer learning tasks on HGs.

## [Conditional Diffusion Process for Inverse Halftoning](#)

- Hao Jiang · Yadong Mu
- abstract@[open-review](#): Inverse halftoning is a technique used to recover realistic images from ancient prints (e.g., photographs, newspapers, books). The rise of deep learning has led to the gradual incorporation of neural network designs into inverse halftoning methods. Most of existing inverse halftoning methods adopt the U-net architecture, which uses an encoder to encode halftone prints, followed by a decoder for image reconstruction. However, the mainstream U-net architecture has poor generalization ability in practical applications. Specifically, when there is a large gap between the halftone patterns of the training and test data, the reconstructed continuous-tone images have obvious artifacts. This is an important issue in practical applications, since the algorithms for generating halftones are ever-evolving. Even for the same algorithm, different parameter choices will result in different halftone dithering patterns. In this paper, we propose the first generative halftoning method in the literature, which regards the black pixels in halftones as physically moving particles, and makes the randomly distributed particles move under some certain guidance through the reverse diffusion process, so as to obtain the desired halftone patterns. In particular, we propose a Conditional Diffusion model for image Halftoning (CDH), which consists of a halftone dithering process and an inverse halftoning process. By changing the initial state of the diffusion model, our method can generate visually plausible halftones with different dithering patterns under the condition of image gray level and Laplacian prior. To avoid introducing redundant patterns and undesired artifacts, we propose a meta-halftone guided network to incorporate blue noise guidance in the diffusion process. In this way, halftone images subject to more diverse distributions are fed into the inverse halftoning model, which helps the model to learn a more robust mapping from the halftone distribution to continuous-tone distribution, thereby improving the generalization ability to unseen samples. Quantitative and qualitative experimental results demonstrate that the proposed method achieves state-of-the-art results.

## [Analyzing Lottery Ticket Hypothesis from PAC-Bayesian Theory Perspective](#)

- Keitaro Sakamoto · Issei Sato
- abstract@[open-review](#): The lottery ticket hypothesis (LTH) has attracted attention because it can explain why over-parameterized models often show high generalization ability. It is known that when we use iterative magnitude pruning (IMP), which is an algorithm to find sparse networks with high generalization ability that can be trained from the initial weights independently, called winning tickets, the initial large learning rate does not work well in deep neural networks such as ResNet. However, since the initial large learning rate generally helps the optimizer to converge to flatter minima, we hypothesize that the winning tickets have relatively sharp minima, which is considered a disadvantage in terms of generalization ability. In this paper, we confirm this hypothesis and show that the PAC-Bayesian theory can provide an explicit understanding of the relationship between LTH and generalization behavior. On the basis of our experimental findings that IMP with a small learning rate finds relatively sharp minima and that the distance from the initial weights is deeply involved in winning tickets, we offer the PAC-Bayes bound using a spike-and-slab distribution to analyze winning tickets. Finally, we revisit existing algorithms for finding winning tickets from a PAC-Bayesian perspective and provide new insights into these methods.

## [Non-Stationary Bandits under Recharging Payoffs: Improved Planning with Sublinear Regret](#)

- Orestis Papadigenopoulos · Constantine Caramanis · Sanjay Shakkottai
- abstract@[open-review](#): The stochastic multi-armed bandit setting has been recently studied in the non-stationary regime, where the mean payoff of each action is a non-decreasing function of the number of rounds passed since it was last played. This model captures natural behavioral aspects of the users which crucially determine the performance of recommendation platforms, ad placement systems, and more. Even assuming prior knowledge of the mean payoff functions, computing an optimal planning in the above model is NP-hard, while the state-of-the-art is a \$1/4\$-approximation algorithm for the case where at most one arm can be played per round. We first focus on the setting where the mean payoff functions are known. In this setting, we significantly improve the best-known guarantees for the planning problem by developing a polynomial-time \$(1-\epsilon)/\epsilon\$-approximation algorithm (asymptotically and in expectation), based on a novel combination of randomized LP rounding and a time-correlated (interleaved) scheduling method. Furthermore, our algorithm achieves improved guarantees -- compared to prior work -- for the case where more than one arms can be played at each round. Moving to the bandit setting, when the mean payoff functions are initially unknown, we show how our algorithm can be transformed into a bandit algorithm with sublinear regret.

## [Emergence of Hierarchical Layers in a Single Sheet of Self-Organizing Spiking Neurons](#)

- Paul Bertens · Seong-Whan Lee
- abstract@[open-review](#): Traditionally convolutional neural network architectures have been designed by stacking layers on top of each other to form deeper hierarchical networks. The cortex in the brain however does not just stack layers as done in standard convolution neural networks, instead different regions are organized next to each other in a large single sheet of neurons. Biological neurons self organize to form topographic maps, where neurons encoding similar stimuli group together to form logical clusters. Here we propose new self-organization principles that allow for the formation of hierarchical cortical regions (i.e. layers) in a completely unsupervised manner without requiring any predefined architecture. Synaptic connections are dynamically grown and pruned, which allows us to actively constrain the number of incoming and outgoing connections. This way we can minimize the wiring cost by taking into account both the synaptic strength and the connection length. The proposed method uses purely local learning rules in the form of spike-timing-dependent plasticity (STDP) with lateral excitation and inhibition. We show experimentally that these self-organization rules are sufficient

for topographic maps and hierarchical layers to emerge. Our proposed Self-Organizing Neural Sheet (SONS) model can thus form traditional neural network layers in a completely unsupervised manner from just a single large pool of unstructured spiking neurons.

## [Mix and Reason: Reasoning over Semantic Topology with Data Mixing for Domain Generalization](#)

- Chaoqi Chen · Luyao Tang · Feng Liu · Gangming Zhao · Yue Huang · Yizhou Yu
- abstract@[open-review](#): Domain generalization (DG) enables generalizing a learning machine from multiple seen source domains to an unseen target one. The general objective of DG methods is to learn semantic representations that are independent of domain labels, which is theoretically sound but empirically challenged due to the complex mixture of common and domain-specific factors. Although disentangling the representations into two disjoint parts has been gaining momentum in DG, the strong presumption over the data limits its efficacy in many real-world scenarios. In this paper, we propose Mix and Reason (MiRe), a new DG framework that learns semantic representations via enforcing the structural invariance of semantic topology. MiRe consists of two key components, namely, Category-aware Data Mixing (CDM) and Adaptive Semantic Topology Refinement (ASTR). CDM mixes two images from different domains in virtue of activation maps generated by two complementary classification losses, making the classifier focus on the representations of semantic objects. ASTR introduces relation graphs to represent semantic topology, which is progressively refined via the interactions between local feature aggregation and global cross-domain relational reasoning. Experiments on multiple DG benchmarks validate the effectiveness and robustness of the proposed MiRe.

## [Modeling Neural Population Activity with Spatiotemporal Transformer](#)

- Trung Le · Eli Shlizerman
- abstract@[open-review](#): Modeling neural population dynamics underlying noisy single-trial spiking activities is essential for relating neural observation and behavior. A recent non-recurrent method - Neural Data Transformers (NDT) - has shown great success in capturing neural dynamics with low inference latency without an explicit dynamical model. However, NDT focuses on modeling the temporal evolution of the population activity while neglecting the rich covariation between individual neurons. In this paper we introduce SpatioTemporal Neural Data Transformer (STNDT), an NDT-based architecture that explicitly models responses of individual neurons in the population across time and space to uncover their underlying firing rates. In addition, we propose a contrastive learning loss that works in accordance with mask modeling objective to further improve the predictive performance. We show that our model achieves state-of-the-art performance on ensemble level in estimating neural activities across four neural datasets, demonstrating its capability to capture autonomous and non-autonomous dynamics spanning different cortical regions while being completely agnostic to the specific behaviors at hand. Furthermore, STNDT spatial attention mechanism reveals consistently important subsets of neurons that play a vital role in driving the response of the entire population, providing interpretability and key insights into how the population of neurons performs computation.

## [Double Bubble, Toil and Trouble: Enhancing Certified Robustness through Transitivity](#)

- Andrew Cullen · Paul Montague · Shijie Liu · Sarah Erfani · Benjamin Rubinstein
- abstract@[open-review](#): The sensitivity of Neural Networks to adversarial attacks can be defrayed through the implementation of specific defences. However, these defences are inherently tied to a single attack (or class of attacks), and as such there is always the potential for a motivated attacker to evade the deployed defence. Recent research has responded to this concern by emphasising the potential for Certified Robustness to provide guarantees of resistance to all norm bounded attacks. While proofs exist demonstrating best-possible robustness for  $\|L_2\|$ -norm bounded attacks, these certificates still are lower bounds on the distance between a point of interest and its nearest adversarial example. In this work, we demonstrate how these best-possible certificates can be improved upon by exploiting both the transitivity of certifications, and the geometry of the input space, giving rise to what we have called Geometrically Informed Certified Robustness. Incorporating these improvements leads to an improvement in the certification for more than 80% of samples in experiments against Tiny-Imagenet, with an on average 5% increase in the associated radius of certification.

## [Max-Min Off-Policy Actor-Critic Method Focusing on Worst-Case Robustness to Model Misspecification](#)

- Takumi Tanabe · Rei Sato · Kazuto Fukuchi · Jun Sakuma · Youhei Akimoto
- abstract@[open-review](#): In the field of reinforcement learning, owing to the high cost and risk of policy training in the real world, policies trained in a simulation environment are often transferred corresponding real-world environment. However, the simulation environment does not perfectly mimic the real-world environment, leading to model misspecification occurs. Multiple studies report significant deterioration of policy performance in a real-world environment. In this study, we focus on scenarios involving a simulation environment with uncertainty parameters and the set of their possible values, called the uncertainty parameter set. The aim is to optimize the worst-case performance on the uncertainty parameter set to guarantee the performance in the corresponding real-world environment, if it is included in the uncertainty parameter set. To obtain a policy that optimizes the worst-case performance, we propose an off-policy actor-critic approach called the Max-Min Twin Delayed Deep Deterministic Policy Gradient Algorithm (M2TD3), which solves a max-min optimization problem using a simultaneous gradient ascent descent approach. Experiments in Multi-Joint Dynamics with Contact (MuJoCo) environments show that the proposed method exhibited a worst-case performance superior to several baseline approaches.

## [Iterative Feature Matching: Toward Provable Domain Generalization with Logarithmic Environments](#)

- Yining Chen · Elan Rosenfeld · Mark Sellke · Tengyu Ma · Andrej Risteski
- abstract@[open-review](#): Domain generalization aims at performing well on unseen test environments with data from a limited number of training environments. Despite a proliferation of proposed algorithms for this task, assessing their performance both theoretically and empirically is still very challenging. Distributional matching algorithms such as (Conditional) Domain Adversarial Networks [Ganin et al., 2016, Long et al., 2018] are popular and enjoy empirical success, but they lack formal guarantees. Other approaches such as Invariant Risk Minimization (IRM) require a prohibitively large number of training environments---linear in the dimension of the spurious feature space  $\|d_s\|$ ---even on simple data models like the one proposed by [Rosenfeld et al., 2021]. Under a variant of this model, we show that ERM and IRM can fail to find the optimal invariant predictor with  $\mathcal{O}(d_s)$  environments. We then present an iterative feature matching algorithm that is guaranteed with high probability to find the optimal invariant predictor after seeing only  $\mathcal{O}(\log d_s)$  environments. Our results provide the first theoretical justification for distribution-matching algorithms widely used in practice under a concrete nontrivial data model.

## [Few-shot Task-agnostic Neural Architecture Search for Distilling Large Language Models](#)

- Dongkuan (DK) Xu · Subhabrata Mukherjee · Xiaodong Liu · Debadeepa Dey · Wenhui Wang · Xiang Zhang · Ahmed Awadallah · Jianfeng Gao
- abstract@[open-review](#): Traditional knowledge distillation (KD) methods manually design student architectures to compress large models given pre-specified computational cost. This requires several trials to find viable students, and repeating the process with change in computational budget. We use Neural Architecture Search (NAS) to automatically distill several compressed students with variable cost from a large model. Existing NAS methods train a single SuperLM consisting of millions of subnetworks with weight-sharing, resulting in interference between subnetworks of different sizes. Additionally, many of these works are task-specific requiring task labels for SuperLM training. Our framework AutoDistil addresses above challenges with the following steps: (a) Incorporates inductive bias and heuristics to partition Transformer search space into K compact sub-spaces (e.g., K=3 can generate typical student sizes of base, small and tiny); (b) Trains one SuperLM for each sub-space using task-agnostic objective (e.g., self-attention distillation) with weight-sharing of students; (c) Lightweight search for the optimal student without re-training. Task-agnostic training and search allow

students to be reused for fine-tuning on any downstream task. Experiments on GLUE benchmark demonstrate AutoDistil to outperform state-of-the-art KD and NAS methods with upto 3x reduction in computational cost and negligible loss in task performance.

## [PAC: Assisted Value Factorisation with Counterfactual Predictions in Multi-Agent Reinforcement Learning](#)

- Hanhan Zhou · Tian Lan · Vaneet Aggarwal
- abstract@[open-review](#): Multi-agent reinforcement learning (MARL) has witnessed significant progress with the development of value function factorization methods. It allows optimizing a joint action-value function through the maximization of factorized per-agent utilities due. In this paper, we show that in partially observable MARL problems, an agent's ordering over its own actions could impose concurrent constraints (across different states) on the representable function class, causing significant estimation error during training. We tackle this limitation and propose PAC, a new framework leveraging Assistive information generated from Counterfactual Predictions of optimal joint action selection, which enable explicit assistance to value function factorization through a novel counterfactual loss. A variational inference-based information encoding method is developed to collect and encode the counterfactual predictions from an estimated baseline. To enable decentralized execution, we also derive factorized per-agent policies inspired by a maximum-entropy MARL framework. We evaluate the proposed PAC on multi-agent predator-prey and a set of StarCraft II micromanagement tasks. Empirical results demonstrate improved results of PAC over state-of-the-art value-based and policy-based multi-agent reinforcement learning algorithms on all benchmarks.

## [Oscillatory Tracking of Continuous Attractor Neural Networks Account for Phase Precession and Procession of Hippocampal Place Cells](#)

- Tianhao Chu · Zilong Ji · Junfeng Zuo · Wenhao Zhang · Tiejun Huang · Yuanyuan Mi · Si Wu
- abstract@[open-review](#): Hippocampal place cells of freely moving rodents display an intriguing temporal organization in their responses known as 'theta phase precession', in which individual neurons fire at progressively earlier phases in successive theta cycles as the animal traverses the place fields. Recent experimental studies found that in addition to phase precession, many place cells also exhibit accompanied phase procession, but the underlying neural mechanism remains unclear. Here, we propose a neural circuit model to elucidate the generation of both kinds of phase shift in place cells' firing. Specifically, we consider a continuous attractor neural network (CANN) with feedback inhibition, which is inspired by the reciprocal interaction between the hippocampus and the medial septum. The feedback inhibition induces intrinsic mobility of the CANN which competes with the extrinsic mobility arising from the external drive. Their interplay generates an oscillatory tracking state, that is, the network bump state (resembling the decoded virtual position of the animal) sweeps back and forth around the external moving input (resembling the physical position of the animal). We show that this oscillatory tracking naturally explains the forward and backward sweeps of the decoded position during the animal's locomotion. At the single neuron level, the forward and backward sweeps account for, respectively, theta phase precession and procession. Furthermore, by tuning the feedback inhibition strength, we also explain the emergence of bimodal cells and unimodal cells, with the former having co-existed phase precession and procession, and the latter having only significant phase precession. We hope that this study facilitates our understanding of hippocampal temporal coding and lays foundation for unveiling their computational functions.

## [Learning Symmetric Rules with SATNet](#)

- SANGHO LIM · Eunyeol Oh · Hongseok Yang
- abstract@[open-review](#): SATNet is a differentiable constraint solver with a custom backpropagation algorithm, which can be used as a layer in a deep-learning system. It is a promising proposal for bridging deep learning and logical reasoning. In fact, SATNet has been successfully applied to learn, among others, the rules of a complex logical puzzle, such as Sudoku, just from input and output pairs where inputs are given as images. In this paper, we show how to improve the learning of SATNet by exploiting symmetries in the target rules of a given but unknown logical puzzle or more generally a logical formula. We present SymSATNet, a variant of SATNet that translates the given symmetries of the target rules to a condition on the parameters of SATNet and requires that the parameters should have a particular parametric form that guarantees the condition. The requirement dramatically reduces the number of parameters to learn for the rules with enough symmetries, and makes the parameter learning of SymSATNet much easier than that of SATNet. We also describe a technique for automatically discovering symmetries of the target rules from examples. Our experiments with Sudoku and Rubik's cube show the substantial improvement of SymSATNet over the baseline SATNet.

## [A Variational Edge Partition Model for Supervised Graph Representation Learning](#)

- Yilin He · Chaojie Wang · Hao Zhang · Bo Chen · Mingyuan Zhou
- abstract@[open-review](#): Graph neural networks (GNNs), which propagate the node features through the edges and learn how to transform the aggregated features under label supervision, have achieved great success in supervised feature extraction for both node-level and graph-level classification tasks. However, GNNs typically treat the graph structure as given and ignore how the edges are formed. This paper introduces a graph generative process to model how the observed edges are generated by aggregating the node interactions over a set of overlapping node communities, each of which contributes to the edges via a logical OR mechanism. Based on this generative model, we partition each edge into the summation of multiple community-specific weighted edges and use them to define community-specific GNNs. A variational inference framework is proposed to jointly learn a GNN based inference network that partitions the edges into different communities, these community-specific GNNs, and a GNN based predictor that combines community-specific GNNs for the end classification task. Extensive evaluations on real-world graph datasets have verified the effectiveness of the proposed method in learning discriminative representations for both node-level and graph-level classification tasks.

## [Transition to Linearity of General Neural Networks with Directed Acyclic Graph Architecture](#)

- Libin Zhu · Chaoyue Liu · Misha Belkin
- abstract@[open-review](#): In this paper we show that feedforward neural networks corresponding to arbitrary directed acyclic graphs undergo transition to linearity as their "width" approaches infinity. The width of these general networks is characterized by the minimum in-degree of their neurons, except for the input and first layers. Our results identify the mathematical structure underlying transition to linearity and generalize a number of recent works aimed at characterizing transition to linearity or constancy of the Neural Tangent Kernel for standard architectures.

## [Diffusion-LM Improves Controllable Text Generation](#)

- Xiang Li · John Thickstun · Ishaan Gulrajani · Percy Liang · Tatsunori Hashimoto
- abstract@[open-review](#): Controlling the behavior of language models (LMs) without re-training is a major open problem in natural language generation. While recent works have demonstrated successes on controlling simple sentence attributes (e.g., sentiment), there has been little progress on complex, fine-grained controls (e.g., syntactic structure). To address this challenge, we develop a new non-autoregressive language model based on continuous diffusions that we call Diffusion-LM. Building upon the recent successes of diffusion models in continuous domains, Diffusion-LM iteratively denoises a sequence of Gaussian vectors into word vectors, yielding a sequence of intermediate latent variables. To control its generation, we iteratively perform gradient updates on these intermediate variables. Diffusion-LM has three properties that enable complex, fine-grained controllable text generation: the continuous nature of diffusion models enables gradient-based control; the non-autoregressive generation order enables more complex, global controls; and incremental denoising induces a coarse-to-fine hierarchy, which facilitates control at multiple granularities. We demonstrate successful control of Diffusion-LM for six challenging fine-grained control tasks, significantly outperforming prior work.

## [Is Sortition Both Representative and Fair?](#)

- Soroush Ebadian · Gregory Kehne · Evi Micha · Ariel Procaccia · Nisarg Shah
- abstract@[open-review](#): Sortition is a form of democracy built on random selection of representatives. Two of the key arguments in favor of sortition are representation (a random panel reflects the composition of the population) and fairness (everyone has a chance to participate). Uniformly random selection is perfectly fair, but is it representative? To answer this question, we introduce the notion of a representation metric on the space of individuals, and assume that the cost of an individual for a panel is determined by the \$q\$-th closest representative; the representation of a (random) panel is measured through the ratio between its (expected) sum of costs over individuals and that of the optimal panel. For \$k/2

## [SoftCore: Unsupervised Anomaly Detection with Noisy Data](#)

- Xi Jiang · Jianlin Liu · Jinbao Wang · Qiang Nie · Kai Wu · Yong Liu · Chengjie Wang · Feng Zheng
- abstract@[open-review](#): Although unsupervised anomaly detection(AD) algorithms perform well in academic datasets, their performance is limited in practical application due to the ideal experimental setting of clean training data. Training with noisy data is an inevitable problem in real-world anomaly detection but is seldom discussed. This paper considers label-level noise in sensory anomaly detection for the first time. To solve this problem, we proposed a memory-based unsupervised AD method, SoftCore, which efficiently denoises the data at the patch level. Noise discriminators are utilized to generate outlier scores for patch-level noise elimination before coresnet construction. The scores are then stored in the memory bank to soften the anomaly detection boundary. Compared with past methods, SoftCore maintains a strong modeling ability of normal data and alleviates the overconfidence problem in coresnet. Comprehensive experiments in various noise scenes demonstrate that SoftCore outperforms the state-of-the-art AD methods on MVTec AD benchmark, and is comparable to those methods under the setting without noise.

## [Weighted Distillation with Unlabeled Examples](#)

- Fotis Iliopoulos · Cenk Baykal · Vasilis Kontonis · Gaurav Menghani · Khoa Trinh · Erik Vee
- abstract@[open-review](#): Distillation with unlabeled examples is a popular and powerful method for training deep neural networks in settings where the amount of labeled data is limited: A large "teacher" neural network is trained on the labeled data available, and then it is used to generate labels on an unlabeled dataset (typically much larger in size). These labels are then utilized to train the smaller "student" model which will actually be deployed. Naturally, the success of the approach depends on the quality of the teacher's labels, since the student could be confused if trained on inaccurate data. This paper proposes a principled approach for addressing this issue based on an importance reweighting scheme tailored to the distillation training paradigm. Our method is hyper-parameter free, data-agnostic, and simple to implement. We demonstrate significant improvements on popular academic datasets when compared to conventional distillation with unlabeled examples. We also accompany our results with a theoretical analysis which rigorously justifies the performance of our method in certain settings.

## [Decentralized, Communication- and Coordination-free Learning in Structured Matching Markets](#)

- Chinmay Maheshwari · Eric Mazumdar · Shankar Sastry
- abstract@[open-review](#): We study the problem of online learning in competitive settings in the context of two-sided matching markets. In particular, one side of the market, the agents, must learn about their preferences over the other side, the firms, through repeated interaction while competing with other agents for successful matches. We propose a class of decentralized, communication- and coordination-free algorithms that agents can use to reach to their stable match in structured matching markets. In contrast to prior works, the proposed algorithms make decisions based solely on an agent's own history of play and requires no foreknowledge of the firms' preferences. Our algorithms are constructed by splitting up the statistical problem of learning one's preferences, from noisy observations, from the problem of competing for firms. We show that under realistic structural assumptions on the underlying preferences of the agents and firms, the proposed algorithms incur a regret which grows at most logarithmically in the time horizon. Our results show that, in the case of matching markets, competition need not drastically affect the performance of decentralized, communication and coordination free online learning algorithms.

## [A Theory of PAC Learnability under Transformation Invariances](#)

- Han Shao · Omar Montasser · Avrim Blum
- abstract@[open-review](#): Transformation invariances are present in many real-world problems. For example, image classification is usually invariant to rotation and color transformation: a rotated car in a different color is still identified as a car. Data augmentation, which adds the transformed data into the training set and trains a model on the augmented data, is one commonly used technique to build these invariances into the learning process. However, it is unclear how data augmentation performs theoretically and what the optimal algorithm is in presence of transformation invariances. In this paper, we study PAC learnability under transformation invariances in three settings according to different levels of realizability: (i) A hypothesis fits the augmented data; (ii) A hypothesis fits only the original data and the transformed data lying in the support of the data distribution; (iii) Agnostic case. One interesting observation is that distinguishing between the original data and the transformed data is necessary to achieve optimal accuracy in setting (ii) and (iii), which implies that any algorithm not differentiating between the original and transformed data (including data augmentation) is not optimal. Furthermore, this type of algorithms can even ``harm'' the accuracy. In setting (i), although it is unnecessary to distinguish between the two data sets, data augmentation still does not perform optimally. Due to such a difference, we propose two combinatorial measures characterizing the optimal sample complexity in setting (i) and (ii)(iii) and provide the optimal algorithms.

## [Rethinking the Reverse-engineering of Trojan Triggers](#)

- Zhenting Wang · Kai Mei · Hailun Ding · Juan Zhai · Shiqing Ma
- abstract@[open-review](#): Deep Neural Networks are vulnerable to Trojan (or backdoor) attacks. Reverse-engineering methods can reconstruct the trigger and thus identify affected models. Existing reverse-engineering methods only consider input space constraints, e.g., trigger size in the input space. Expressly, they assume the triggers are static patterns in the input space and fail to detect models with feature-space triggers such as image style transformations. We observe that both input-space and feature-space Trojans are associated with feature space hyperplanes. Based on this observation, we design a novel reverse-engineering method that exploits the feature space constraint to reverse-engineer Trojan triggers. Results on four datasets and seven different attacks demonstrate that our solution effectively defends both input-space and feature-space Trojans. It outperforms state-of-the-art reverse-engineering methods and other types of defenses in both Trojaned model detection and mitigation tasks. On average, the detection accuracy of our method is 93%. For Trojan mitigation, our method can reduce the ASR (attack success rate) to only 0.26% with the BA (benign accuracy) remaining nearly unchanged. Our code can be found in <https://anonymous.4open.science/r/FeatureRE-10B7>.

## [Beyond Not-Forgetting: Continual Learning with Backward Knowledge Transfer](#)

- Sen Lin · Li Yang · Deliang Fan · Junshan Zhang
- abstract@[open-review](#): By learning a sequence of tasks continually, an agent in continual learning (CL) can improve the learning performance of both a new task and `old' tasks by leveraging the forward knowledge transfer and the backward knowledge transfer, respectively. However, most existing CL methods focus on addressing catastrophic forgetting in neural networks by minimizing the modification of the learnt model for old tasks. This inevitably limits the backward knowledge transfer from the new task to the old tasks, because judicious model updates could possibly improve the learning

performance of the old tasks as well. To tackle this problem, we first theoretically analyze the conditions under which updating the learnt model of old tasks could be beneficial for CL and also lead to backward knowledge transfer, based on the gradient projection onto the input subspaces of old tasks. Building on the theoretical analysis, we next develop a ContinUal learning method with Backward knowlEdge tRansfer (CUBER), for a fixed capacity neural network without data replay. In particular, CUBER first characterizes the task correlation to identify the positively correlated old tasks in a layer-wise manner, and then selectively modifies the learnt model of the old tasks when learning the new task. Experimental studies show that CUBER can even achieve positive backward knowledge transfer on several existing CL benchmarks for the first time without data replay, where the related baselines still suffer from catastrophic forgetting (negative backward knowledge transfer). The superior performance of CUBER on the backward knowledge transfer also leads to higher accuracy accordingly.

## [Large Language Models are Zero-Shot Reasoners](#)

- Takeshi Kojima · Shixiang (Shane) Gu · Machel Reid · Yutaka Matsuo · Yusuke Iwasawa
- abstract@[open-review](#): Pretrained large language models (LLMs) are widely used in many sub-fields of natural language processing (NLP) and generally known as excellent few-shot learners with task-specific exemplars. Notably, chain of thought (CoT) prompting, a recent technique for eliciting complex multi-step reasoning through step-by-step answer examples, achieved the state-of-the-art performances in arithmetics and symbolic reasoning, difficult system-2 tasks that do not follow the standard scaling laws for LLMs. While these successes are often attributed to LLMs' ability for few-shot learning, we show that LLMs are decent zero-shot reasoners by simply adding ``Let's think step by step'' before each answer. Experimental results demonstrate that our Zero-shot-CoT, using the same single prompt template, significantly outperforms zero-shot LLM performances on diverse benchmark reasoning tasks including arithmetics (MultiArith, GSM8K, AQUA-RAT, SVAMP), symbolic reasoning (Last Letter, Coin Flip), and other logical reasoning tasks (Date Understanding, Tracking Shuffled Objects), without any hand-crafted few-shot examples, e.g. increasing the accuracy on MultiArith from 17.7% to 78.7% and GSM8K from 10.4% to 40.7% with 175B parameter InstructGPT model, as well as similar magnitudes of improvements with another off-the-shelf large model, 540B parameter PaLM. The versatility of this single prompt across very diverse reasoning tasks hints at untapped and understudied fundamental zero-shot capabilities of LLMs, suggesting high-level, multi-task broad cognitive capabilities may be extracted by simple prompting. We hope our work not only serves as the minimal strongest zero-shot baseline for the challenging reasoning benchmarks, but also highlights the importance of carefully exploring and analyzing the enormous zero-shot knowledge hidden inside LLMs before crafting finetuning datasets or few-shot exemplars.

## [Offline Goal-Conditioned Reinforcement Learning via \\$f\\$-Advantage Regression](#)

- Jason Yecheng Ma · Jason Yan · Dinesh Jayaraman · Osbert Bastani
- abstract@[open-review](#): Offline goal-conditioned reinforcement learning (GCRL) promises general-purpose skill learning in the form of reaching diverse goals from purely offline datasets. We propose \$\\textbf{\\{Go\\}}\$al-conditioned \$f\$-\$\\textbf{\\{A\\}}\$dvantage \$\\textbf{\\{R\\}}\$egression (GoFAR), a novel regression-based offline GCRL algorithm derived from a state-occupancy matching perspective; the key intuition is that the goal-reaching task can be formulated as a state-occupancy matching problem between a dynamics-abiding imitator agent and an expert agent that directly teleports to the goal. In contrast to prior approaches, GoFAR does not require any hindsight relabeling and enjoys uninterleaved optimization for its value and policy networks. These distinct features confer GoFAR with much better offline performance and stability as well as statistical performance guarantee that is unattainable for prior methods. Furthermore, we demonstrate that GoFAR's training objectives can be re-purposed to learn an agent-independent goal-conditioned planner from purely offline source-domain data, which enables zero-shot transfer to new target domains. Through extensive experiments, we validate GoFAR's effectiveness in various problem settings and tasks, significantly outperforming prior state-of-art. Notably, on a real robotic dexterous manipulation task, while no other method makes meaningful progress, GoFAR acquires complex manipulation behavior that successfully accomplishes diverse goals.

## [Pre-Trained Model Reusability Evaluation for Small-Data Transfer Learning](#)

- Yao-Xiang Ding · Xi-Zhu Wu · Kun Zhou · Zhi-Hua Zhou
- abstract@[open-review](#): We study {it model reusability evaluation} (MRE) for source pre-trained models: evaluating their transfer learning performance to new target tasks. In special, we focus on the setting under which the target training datasets are small, making it difficult to produce reliable MRE scores using them. Under this situation, we propose {it synergistic learning} for building the task-model metric, which can be realized by collecting a set of pre-trained models and asking a group of data providers to participate. We provide theoretical guarantees to show that the learned task-model metric distances can serve as trustworthy MRE scores, and propose synergistic learning algorithms and models for general learning tasks. Experiments show that the MRE models learned by synergistic learning can generate significantly more reliable MRE scores than existing approaches for small-data transfer learning.

## [Effects of Data Geometry in Early Deep Learning](#)

- Saket Tiwari · George Konidaris
- abstract@[open-review](#): Deep neural networks can approximate functions on different types of data, from images to graphs, with varied underlying structure. This underlying structure can be viewed as the geometry of the data manifold. By extending recent advances in the theoretical understanding of neural networks, we study how a randomly initialized neural network with piecewise linear activation splits the data manifold into regions where the neural network behaves as a linear function. We derive bounds on the density of boundary of linear regions and the distance to these boundaries on the data manifold. This leads to insights into the expressivity of randomly initialized deep neural networks on non-Euclidean data sets. We empirically corroborate our theoretical results using a toy supervised learning problem. Our experiments demonstrate that number of linear regions varies across manifolds and the results hold with changing neural network architectures. We further demonstrate how the complexity of linear regions is different on the low dimensional manifold of images as compared to the Euclidean space, using the MetFaces dataset.

## [Asymptotic Behaviors of Projected Stochastic Approximation: A Jump Diffusion Perspective](#)

- Jiadong Liang · Yuze Han · Xiang Li · Zhihua Zhang
- abstract@[open-review](#): In this paper, we consider linearly constrained stochastic approximation problems with federated learning (FL) as a special case. We propose a stochastic approximation algorithm named by LPSA with probabilistic projections to ensure feasibility so that projections are performed with probability \$p\_n\$ at the \$n\$-th iteration. Considering a specific family of the probability \$p\_n\$ and step size \$\eta\_n\$, we analyze our algorithm from an asymptotic and continuous perspective. Using a novel jump diffusion approximation, we show that the trajectories consisting of properly rescaled last iterates weakly converge to the solution of specific SDEs. By analyzing the SDEs, we identify the asymptotic behaviors of LPSA for different choices of \$(p\_n, \eta\_n)\$. We find the algorithm presents an intriguing asymptotic bias-variance trade-off according to the relative magnitude of \$p\_n\$ w.r.t. \$\eta\_n\$. It provides insights on how to choose appropriate \$\{(p\_n, \eta\_n)\}\_{n \geq 1}\$ to minimize the projection complexity.

## [Using Embeddings for Causal Estimation of Peer Influence in Social Networks](#)

- Irina Cristali · Victor Veitch
- abstract@[open-review](#): We address the problem of using observational data to estimate peer contagion effects, the influence of treatments applied to individuals in a network on the outcomes of their neighbors. A main challenge to such estimation is that homophily - the tendency of connected units to share similar latent traits - acts as an unobserved confounder for contagion effects. Informally, it's hard to tell whether your friends have similar outcomes because they were influenced by your treatment, or whether it's due to some common trait that caused you to be friends in the first place. Because these

common causes are not usually directly observed, they cannot be simply adjusted for. We describe an approach to perform the required adjustment using node embeddings learned from the network itself. The main aim is to perform this adjustment nonparametrically, without functional form assumptions on either the process that generated the network or the treatment assignment and outcome processes. The key contributions are to nonparametrically formalize the causal effect in a way that accounts for homophily, and to show how embedding methods can be used to identify and estimate this effect.

## [Bandit Theory and Thompson Sampling-Guided Directed Evolution for Sequence Optimization](#)

- Hui Yuan · Chengzhuo Ni · Huazheng Wang · Xuezhou Zhang · Le Cong · Csaba Szepesvari · Mengdi Wang
- abstract@[open-review](#): Directed Evolution (DE), a landmark wet-lab method originated in 1960s, enables discovery of novel protein designs via evolving a population of candidate sequences. Recent advances in biotechnology has made it possible to collect high-throughput data, allowing the use of machine learning to map out a protein's sequence-to-function relation. There is a growing interest in machine learning-assisted DE for accelerating protein optimization. Yet the theoretical understanding of DE, as well as the use of machine learning in DE, remains limited. In this paper, we connect DE with the bandit learning theory and make a first attempt to study regret minimization in DE. We propose a Thompson Sampling-guided Directed Evolution (TS-DE) framework for sequence optimization, where the sequence-to-function mapping is unknown and querying a single value is subject to costly and noisy measurements. TS-DE updates a posterior of the function based on collected measurements. It uses a posterior-sampled function estimate to guide the crossover recombination and mutation steps in DE. In the case of a linear model, we show that TS-DE enjoys a Bayesian regret of order  $\tilde{O}(d^2 \sqrt{MT})$ , where  $d$  is feature dimension,  $M$  is population size and  $T$  is number of rounds. This regret bound is nearly optimal, confirming that bandit learning can provably accelerate DE. It may have implications for more general sequence optimization and evolutionary algorithms.

## [Kernel similarity matching with Hebbian networks](#)

- Kyle Luther · Sebastian Seung
- abstract@[open-review](#): Recent works have derived neural networks with online correlation-based learning rules to perform \textit{kernel similarity matching}. These works applied existing linear similarity matching algorithms to nonlinear features generated with random Fourier methods. In this paper attempt to perform kernel similarity matching by directly learning the nonlinear features. Our algorithm proceeds by deriving and then minimizing an upper bound for the sum of squared errors between output and input kernel similarities. The construction of our upper bound leads to online correlation-based learning rules which can be implemented with a 1 layer recurrent neural network. In addition to generating high-dimensional linearly separable representations, we show that our upper bound naturally yields representations which are sparse and selective for specific input patterns. We compare the approximation quality of our method to neural random Fourier method and variants of the popular but non-biological ``Nyström'' method for approximating the kernel matrix. Our method appears to be comparable or better than randomly sampled Nyström methods when the outputs are relatively low dimensional (although still potentially higher dimensional than the inputs) but less faithful when the outputs are very high dimensional.

## [Your Out-of-Distribution Detection Method is not Robust!](#)

- Mohammad Azizmalayeri · Arshia Soltani Moakhar · Arman Zarei · Reihaneh Zohrabi · Mohammad Manzuri · Mohammad Hossein Rohban
- abstract@[open-review](#): Out-of-distribution (OOD) detection has recently gained substantial attention due to the importance of identifying out-of-domain samples in reliability and safety. Although OOD detection methods have advanced by a great deal, they are still susceptible to adversarial examples, which is a violation of their purpose. To mitigate this issue, several defenses have recently been proposed. Nevertheless, these efforts remained ineffective, as their evaluations are based on either small perturbation sizes, or weak attacks. In this work, we re-examine these defenses against an end-to-end PGD attack on in/out data with larger perturbation sizes, e.g. up to commonly used  $\epsilon=8/255$  for the CIFAR-10 dataset. Surprisingly, almost all of these defenses perform worse than a random detection under the adversarial setting. Next, we aim to provide a robust OOD detection method. In an ideal defense, the training should expose the model to almost {it all} possible adversarial perturbations, which can be achieved through adversarial training. That is, such training perturbations should be based on both in- and out-of-distribution samples. Therefore, unlike OOD detection in the standard setting, access to OOD, as well as in-distribution, samples sounds necessary in the adversarial training setup. These tips lead us to adopt generative OOD detection methods, such as OpenGAN, as a baseline. We subsequently propose the Adversarially Trained Discriminator (ATD), which utilizes a pre-trained robust model to extract robust features, and a generator model to create OOD samples. We noted that, for the sake of training stability, in the adversarial training of the discriminator, one should attack real in-distribution as well as real outliers, but not generated outliers. Using ATD with CIFAR-10 and CIFAR-100 as the in-distribution data, we could significantly outperform all previous methods in the robust AUROC while maintaining high standard AUROC and classification accuracy.

## [On Neural Network Pruning's Effect on Generalization](#)

- Tian Jin · Daniel M Roy · Michael Carbin · Jonathan Frankle · Gintare Karolina Dziugaite
- abstract@[open-review](#): Practitioners frequently observe that pruning improves model generalization. A long-standing hypothesis attributes such improvement to model size reduction. However, recent studies on over-parameterization characterize a new model size regime, in which larger models achieve better generalization. A contradiction arises when pruning is applied to over-parameterized models -- while theory predicts that reducing size harms generalization, pruning nonetheless improves it. Motivated by such a contradiction, we re-examine pruning's effect on generalization empirically. We demonstrate that pruning's generalization-improving effect cannot be fully accounted for by weight removal. Instead, we find that pruning can lead to better optimization, improving training loss. We find that pruning can also lead to stronger regularization, mitigating the harmful effect of noisy examples. We empirically demonstrate that better optimization through extending training time or stronger regularization through reducing model width alone cannot explain the full extent of pruning's generalization-improving effects.

## [Calibrated Data-Dependent Constraints with Exact Satisfaction Guarantees](#)

- Songkai Xue · Yuekai Sun · Mikhail Yurochkin
- abstract@[open-review](#): We consider the task of training machine learning models with data-dependent constraints. Such constraints often arise as empirical versions of expected value constraints that enforce fairness or stability goals. We reformulate data-dependent constraints so that they are calibrated: enforcing the reformulated constraints guarantees that their expected value counterparts are satisfied with a user-prescribed probability. The resulting optimization problem is amendable to standard stochastic optimization algorithms, and we demonstrate the efficacy of our method on a fairness-sensitive classification task where we wish to guarantee the classifier's fairness (at test time).

## [On the Sample Complexity of Stabilizing LTI Systems on a Single Trajectory](#)

- Yang Hu · Adam Wierman · Guannan Qu
- abstract@[open-review](#): Stabilizing an unknown dynamical system is one of the central problems in control theory. In this paper, we study the sample complexity of the learn-to-stabilize problem in Linear Time-Invariant (LTI) systems on a single trajectory. Current state-of-the-art approaches require a sample complexity linear in  $n$ , the state dimension, which incurs a state norm that blows up exponentially in  $n$ . We propose a novel algorithm based on spectral decomposition that only needs to learn ``a small part'' of the dynamical matrix acting on its unstable subspace. We show that, under proper assumptions, our algorithm stabilizes an LTI system on a single trajectory with  $\tilde{O}(k)$  samples, where  $k$  is the instability index of the system. This represents the first sub-linear sample complexity result for the stabilization of LTI systems under the regime when  $k = o(n)$ .

## [Dataset Distillation using Neural Feature Regression](#)

- Yongchao Zhou · Ehsan Nezhadarya · Jimmy Ba
- abstract@[open-review](#): Dataset distillation aims to learn a small synthetic dataset that preserves most of the information from the original dataset. Dataset distillation can be formulated as a bi-level meta-learning problem where the outer loop optimizes the meta-dataset and the inner loop trains a model on the distilled data. Meta-gradient computation is one of the key challenges in this formulation, as differentiating through the inner loop learning procedure introduces significant computation and memory costs. In this paper, we address these challenges using neural Feature Regression with Pooling (FRePo), achieving the state-of-the-art performance with an order of magnitude less memory requirement and two orders of magnitude faster training than previous methods. The proposed algorithm is analogous to truncated backpropagation through time with a pool of models to alleviate various types of overfitting in dataset distillation. FRePo significantly outperforms the previous methods on CIFAR100, Tiny ImageNet, and ImageNet-1K. Furthermore, we show that high-quality distilled data can greatly improve various downstream applications, such as continual learning and membership inference defense.

## [Accelerated Training of Physics Informed Neural Networks \(PINNs\) using Meshless Discretizations](#)

- Ramansh Sharma · Varun Shankar
- abstract@[open-review](#): Physics-informed neural networks (PINNs) are neural networks trained by using physical laws in the form of partial differential equations (PDEs) as soft constraints. We present a new technique for the accelerated training of PINNs that combines modern scientific computing techniques with machine learning: discretely-trained PINNs (DT-PINNs). The repeated computation of the partial derivative terms in the PINN loss functions via automatic differentiation during training is known to be computationally expensive, especially for higher-order derivatives. DT-PINNs are trained by replacing these exact spatial derivatives with high-order accurate numerical discretizations computed using meshless radial basis function-finite differences (RBF-FD) and applied via sparse-matrix vector multiplication. While in principle any high-order discretization may be used, the use of RBF-FD allows for DT-PINNs to be trained even on point cloud samples placed on irregular domain geometries. Additionally, though traditional PINNs (vanilla-PINNs) are typically stored and trained in 32-bit floating-point (fp32) on the GPU, we show that for DT-PINNs, using fp64 on the GPU leads to significantly faster training times than fp32 vanilla-PINNs with comparable accuracy. We demonstrate the efficiency and accuracy of DT-PINNs via a series of experiments. First, we explore the effect of network depth on both numerical and automatic differentiation of a neural network with random weights and show that RBF-FD approximations of third-order accuracy and above are more efficient while being sufficiently accurate. We then compare the DT-PINNs to vanilla-PINNs on both linear and nonlinear Poisson equations and show that DT-PINNs achieve similar losses with 2-4x faster training times on a consumer GPU. Finally, we also demonstrate that similar results can be obtained for the PINN solution to the heat equation (a space-time problem) by discretizing the spatial derivatives using RBF-FD and using automatic differentiation for the temporal derivative. Our results show that fp64 DT-PINNs offer a superior cost-accuracy profile to fp32 vanilla-PINNs, opening the door to a new paradigm of leveraging scientific computing techniques to support machine learning.

## [Conformal Prediction with Temporal Quantile Adjustments](#)

- Zhen Lin · Shubhendu Trivedi · Jimeng Sun
- abstract@[open-review](#): We develop Temporal Quantile Adjustment (TQA), a general method to construct efficient and valid prediction intervals (PIs) for regression on cross-sectional time series data. Such data is common in many domains, including econometrics and healthcare. A canonical example in healthcare is predicting patient outcomes using physiological time-series data, where a population of patients composes a cross-section. Reliable PI estimators in this setting must address two distinct notions of coverage: cross-sectional coverage across a cross-sectional slice, and longitudinal coverage along the temporal dimension for each time series. Recent works have explored adapting Conformal Prediction (CP) to obtain PIs in the time series context. However, none handles both notions of coverage simultaneously. CP methods typically query a pre-specified quantile from the distribution of nonconformity scores on a calibration set. TQA adjusts the quantile to query in CP at each time  $\$$ , accounting for both cross-sectional and longitudinal coverage in a theoretically-grounded manner. The post-hoc nature of TQA facilitates its use as a general wrapper around any time series regression model. We validate TQA's performance through extensive experimentation: TQA generally obtains efficient PIs and improves longitudinal coverage while preserving cross-sectional coverage.

## [Domain Adaptation meets Individual Fairness. And they get along.](#)

- Debarghya Mukherjee · Felix Petersen · Mikhail Yurochkin · Yuekai Sun
- abstract@[open-review](#): Many instances of algorithmic bias are caused by distributional shifts. For example, machine learning (ML) models often perform worse on demographic groups that are underrepresented in the training data. In this paper, we leverage this connection between algorithmic fairness and distribution shifts to show that algorithmic fairness interventions can help ML models overcome distribution shifts, and that domain adaptation methods (for overcoming distribution shifts) can mitigate algorithmic biases. In particular, we show that (i) enforcing suitable notions of individual fairness (IF) can improve the out-of-distribution accuracy of ML models under the covariate shift assumption and that (ii) it is possible to adapt representation alignment methods for domain adaptation to enforce individual fairness. The former is unexpected because IF interventions were not developed with distribution shifts in mind. The latter is also unexpected because representation alignment is not a common approach in the individual fairness literature.

## [Where do Models go Wrong? Parameter-Space Saliency Maps for Explainability](#)

- Roman Levin · Manli Shu · Eitan Borgnia · Furong Huang · Micah Goldblum · Tom Goldstein
- abstract@[open-review](#): Conventional saliency maps highlight input features to which neural network predictions are highly sensitive. We take a different approach to saliency, in which we identify and analyze the network parameters, rather than inputs, which are responsible for erroneous decisions. We first verify that identified salient parameters are indeed responsible for misclassification by showing that turning these parameters off improves predictions on the associated samples more than turning off the same number of random or least salient parameters. We further validate the link between salient parameters and network misclassification errors by observing that fine-tuning a small number of the most salient parameters on a single sample results in error correction on other samples which were misclassified for similar reasons -- nearest neighbors in the saliency space. After validating our parameter-space saliency maps, we demonstrate that samples which cause similar parameters to malfunction are semantically similar. Further, we introduce an input-space saliency counterpart which reveals how image features cause specific network components to malfunction.

## [Score-based generative modeling secretly minimizes the Wasserstein distance](#)

- Dohyun Kwon · Ying Fan · Kangwook Lee
- abstract@[open-review](#): Score-based generative models are shown to achieve remarkable empirical performances in various applications such as image generation and audio synthesis. However, a theoretical understanding of score-based diffusion models is still incomplete. Recently, Song et al. showed that the training objective of score-based generative models is equivalent to minimizing the Kullback-Leibler divergence of the generated distribution from the data distribution. In this work, we show that score-based models also minimize the Wasserstein distance between them. Specifically, we prove that the Wasserstein distance is upper bounded by the square root of the objective function up to multiplicative constants and a fixed constant offset. Our proof is based on a novel application of the theory of optimal transport, which can be of independent interest to the society. Our numerical experiments support our findings. By analyzing our upper bounds, we provide a few techniques to obtain tighter upper bounds.

## [Hedging as Reward Augmentation in Probabilistic Graphical Models](#)

- Debarun Bhattacharjya · Radu Marinescu
- abstract@[open-review](#): Most people associate the term 'hedging' exclusively with financial applications, particularly the use of financial derivatives. We argue that hedging is an activity that human and machine agents should engage in more broadly, even when the agent's value is not necessarily in monetary units. In this paper, we propose a decision-theoretic view of hedging based on augmenting a probabilistic graphical model -- specifically a Bayesian network or an influence diagram -- with a reward. Hedging is therefore posed as a particular kind of graph manipulation, and can be viewed as analogous to control/intervention and information gathering related analysis. Effective hedging occurs when a risk-averse agent finds opportunity to balance uncertain rewards in their current situation. We illustrate the concepts with examples and counter-examples, and conduct experiments to demonstrate the properties and applicability of the proposed computational tools that enable agents to proactively identify potential hedging opportunities in real-world situations.

## [Multiview Human Body Reconstruction from Uncalibrated Cameras](#)

- Zhixuan Yu · Linguang Zhang · Yuanlu Xu · Chengcheng Tang · LUAN TRAN · Cem Keskin · Hyun Soo Park
- abstract@[open-review](#): We present a new method to reconstruct 3D human body pose and shape by fusing visual features from multiview images captured by uncalibrated cameras. Existing multiview approaches often use spatial camera calibration (intrinsic and extrinsic parameters) to geometrically align and fuse visual features. Despite remarkable performances, the requirement of camera calibration restricted their applicability to real-world scenarios, e.g., reconstruction from social videos with wide-baseline cameras. We address this challenge by leveraging the commonly observed human body as a semantic calibration target, which eliminates the requirement of camera calibration. Specifically, we map per-pixel image features to a canonical body surface coordinate system agnostic to views and poses using dense keypoints (correspondences). This feature mapping allows us to semantically, instead of geometrically, align and fuse visual features from multiview images. We learn a self-attention mechanism to reason about the confidence of visual features across and within views. With fused visual features, a regressor is learned to predict the parameters of a body model.

## [Curious Exploration via Structured World Models Yields Zero-Shot Object Manipulation](#)

- Cansu Sancaktar · Sebastian Blaes · Georg Martius
- abstract@[open-review](#): It has been a long-standing dream to design artificial agents that explore their environment efficiently via intrinsic motivation, similar to how children perform curious free play. Despite recent advances in intrinsically motivated reinforcement learning (RL), sample-efficient exploration in object manipulation scenarios remains a significant challenge as most of the relevant information lies in the sparse agent-object and object-object interactions. In this paper, we propose to use structured world models to incorporate relational inductive biases in the control loop to achieve sample-efficient and interaction-rich exploration in compositional multi-object environments. By planning for future novelty inside structured world models, our method generates free-play behavior that starts to interact with objects early on and develops more complex behavior over time. Instead of using models only to compute intrinsic rewards, as commonly done, our method showcases that the self-reinforcing cycle between good models and good exploration also opens up another avenue: zero-shot generalization to downstream tasks via model-based planning. After the entirely intrinsic task-agnostic exploration phase, our method solves challenging downstream tasks such as stacking, flipping, pick & place, and throwing that generalizes to unseen numbers and arrangements of objects without any additional training.

## [An Efficient Framework for Computing Tight Lipschitz Constants of Neural Networks](#)

- Zhouxing Shi · Yihan Wang · Huan Zhang · J. Zico Kolter · Cho-Jui Hsieh
- abstract@[open-review](#): Lipschitz constants are connected to many properties of neural networks, such as robustness, fairness, and generalization. Existing methods for computing Lipschitz constants either are computationally inefficient or produce loose upper bounds. In this paper, we develop an efficient framework for computing the  $\ell_\infty$  local Lipschitz constant of a neural network by tightly upper bounding the norm of Clarke Jacobian. We view the computation for Clarke Jacobian by chain rule as a higher-order backward computational graph. On this graph, we adopt bound-propagation with linear relaxation to obtain provable bounds, and we derive tight linear relaxations for specific nonlinearities in Clarke Jacobian that previous works could not tightly bound. We further tighten our bounds by Branch-and-Bound (BaB) when time budget allows. Experiments show that on tiny models, our method produces comparable bounds compared to exact methods that cannot scale to slightly larger models; on larger models, our method efficiently produces tighter results than existing relaxed or naive methods, and our method scales to much larger practical models that previous works could not handle. We also demonstrate a potential application of our method for provable monotonicity analysis.

## [Dynamic Tensor Product Regression](#)

- Aravind Reddy · Zhao Song · Lichen Zhang
- abstract@[open-review](#): In this work, we initiate the study of Dynamic Tensor Product Regression. One has matrices  $A_1 \in \mathbb{R}^{n_1 \times d_1}, \dots, A_q \in \mathbb{R}^{n_q \times d_q}$  and a label vector  $b \in \mathbb{R}^{n_1 \dots n_q}$ , and the goal is to solve the regression problem with the design matrix  $A$  being the tensor product of the matrices  $A_1, A_2, \dots, A_q$  i.e.  $\min_x \|Ax - b\|_2$ . At each time step, one matrix  $A_i$  receives a sparse change, and the goal is to maintain a sketch of the tensor product  $A_1 \otimes \dots \otimes A_q$  so that the regression solution can be updated quickly. Recomputing the solution from scratch for each round is extremely expensive so it is important to develop algorithms which can quickly update the solution with the new design matrix. Our main result is a dynamic tree data structure where any update to a single matrix can be propagated quickly throughout the tree. We show that our data structure can be used to solve dynamic versions of not only Tensor Product Regression, but also Tensor Product Spline regression (which is a generalization of ridge regression) and for maintaining Low Rank Approximations for the tensor product.

## [Towards Lightweight Black-Box Attack Against Deep Neural Networks](#)

- Chenghao Sun · Yonggang Zhang · Wan Chaoqun · Qizhou Wang · Ya Li · Tongliang Liu · Bo Han · Xinmei Tian
- abstract@[open-review](#): Black-box attacks can generate adversarial examples without accessing the parameters of target model, largely exacerbating the threats of deployed deep neural networks (DNNs). However, previous works state that black-box attacks fail to mislead target models when their training data and outputs are inaccessible. In this work, we argue that black-box attacks can pose practical attacks in this extremely restrictive scenario where only several test samples are available. Specifically, we find that attacking the shallow layers of DNNs trained on a few test samples can generate powerful adversarial examples. As only a few samples are required, we refer to these attacks as lightweight black-box attacks. The main challenge to promoting lightweight attacks is to mitigate the adverse impact caused by the approximation error of shallow layers. As it is hard to mitigate the approximation error with few available samples, we propose Error Transformer (ETF) for lightweight attacks. Namely, ETF transforms the approximation error in the parameter space into a perturbation in the feature space and alleviates the error by disturbing features. In experiments, lightweight black-box attacks with the proposed ETF achieve surprising results. For example, even if only 1 sample per category available, the attack success rate in lightweight black-box attacks is only about 3% lower than that of the black-box attacks with complete training data.

## [Data-Efficient Augmentation for Training Neural Networks](#)

- Tian Yu Liu · Baharan Mirzasoleiman
- abstract@[open-review](#): Data augmentation is essential to achieve state-of-the-art performance in many deep learning applications. However, the most effective augmentation techniques become computationally prohibitive for even medium-sized datasets. To address this, we propose a rigorous technique to select subsets of data points that when augmented, closely capture the training dynamics of full data augmentation. We first show that data

augmentation, modeled as additive perturbations, improves learning and generalization by relatively enlarging and perturbing the smaller singular values of the network Jacobian, while preserving its prominent directions. Then, we propose a framework to iteratively extract small subsets of training data that when augmented, closely capture the alignment of the fully augmented Jacobian with labels/residuals. We prove that stochastic gradient descent applied to augmented subsets found by our approach have similar training dynamics to that of fully augmented data. Our experiments demonstrate that our method outperforms state-of-the-art by 7.7% on CIFAR10 with 6.3x speedup and 4.7% on SVHN with 2.2x speedup, using 10% and 30% augmented subsets respectively. Augmenting 10% and 30% subsets from our method beats random baselines by 7.9% and 5.3% on TinyImageNet, and by 7.6% and 2.3% on ImageNet.

## [On Scrambling Phenomena for Randomly Initialized Recurrent Networks](#)

- Vaggos Chatziafratis · Ioannis Panageas · Clayton Sanford · Stelios Stavroulakis
- abstract@[open-review](#): Recurrent Neural Networks (RNNs) frequently exhibit complicated dynamics, and their sensitivity to the initialization process often renders them notoriously hard to train. Recent works have shed light on such phenomena analyzing when exploding or vanishing gradients may occur, either of which is detrimental for training dynamics. In this paper, we point to a formal connection between RNNs and chaotic dynamical systems and prove a qualitatively stronger phenomenon about RNNs than what exploding gradients seem to suggest. Our main result proves that under standard initialization (e.g., He, Xavier etc.), RNNs will exhibit  $\text{Li-Yorke chaos}$  with  $\text{constant}$  probability  $\text{independent}$  of the network's width. This explains the experimentally observed phenomenon of  $\text{scrambling}$ , under which trajectories of nearby points may appear to be arbitrarily close during some timesteps, yet will be far away in future timesteps. In stark contrast to their feedforward counterparts, we show that chaotic behavior in RNNs is preserved under small perturbations and that their expressive power remains exponential in the number of feedback iterations. Our technical arguments rely on viewing RNNs as random walks under non-linear activations, and studying the existence of certain types of higher-order fixed points called  $\text{periodic points}$  in order to establish phase transitions from order to chaos.

## [Globally Gated Deep Linear Networks](#)

- Qianyi Li · Haim Sompolinsky
- abstract@[open-review](#): Recently proposed Gated Linear Networks (GLNs) present a tractable nonlinear network architecture, and exhibit interesting capabilities such as learning with local error signals and reduced forgetting in sequential learning. In this work, we introduce a novel gating architecture, named Globally Gated Deep Linear Networks (GGDLNs) where gating units are shared among all processing units in each layer, thereby decoupling the architecture of the nonlinear but unlearned gating and the learned linear processing motifs. Using statistical mechanics, we derive exact equations for the generalization properties of Bayesian Learning in these networks in the finite-width thermodynamic limit, defined by  $N, P \rightarrow \infty$  while  $P/N = O(1)$  where  $N$  and  $P$  are the hidden layers' width and size of training data sets, respectively. We find that the statistics of the network predictor can be expressed in terms of kernels that undergo shape renormalization through a data-dependent order-parameter matrix compared to the infinite-width Gaussian Process (GP) kernels. Our theory accurately captures the behavior of finite width GGDLNs trained with gradient descent (GD) dynamics. We show that kernel shape renormalization gives rise to rich generalization properties w.r.t. network width, depth, and  $L_2$  regularization amplitude. Interestingly, networks with a large number of gating units behave similarly to standard ReLU architectures. Although gating units in the model do not participate in supervised learning, we show the utility of unsupervised learning of the gating parameters. Additionally, our theory allows the evaluation of the network capacity for learning multiple tasks by incorporating task-relevant information into the gating units. In summary, our work is the first exact theoretical solution to learning in a family of nonlinear networks with finite width. The rich and diverse behavior of the GGDLNs suggests that they are helpful analytically tractable models of learning single and multiple tasks, in finite-width nonlinear deep networks.

## [Stability and Scalability of Node Perturbation Learning](#)

- Naoki Hiratani · Yash Mehta · Timothy Lillicrap · Peter E Latham
- abstract@[open-review](#): To survive, animals must adapt synaptic weights based on external stimuli and rewards. And they must do so using local, biologically plausible, learning rules -- a highly nontrivial constraint. One possible approach is to perturb neural activity (or use intrinsic, ongoing noise to perturb it), determine whether performance increases or decreases, and use that information to adjust the weights. This algorithm -- known as node perturbation -- has been shown to work on simple problems, but little is known about either its stability or its scalability with respect to network size. We investigate these issues both analytically, in deep linear networks, and numerically, in deep nonlinear ones. We show analytically that in deep linear networks with one hidden layer, both learning time and performance depend very weakly on hidden layer size. However, unlike stochastic gradient descent, when there is model mismatch between the student and teacher networks, node perturbation is always unstable. The instability is triggered by weight diffusion, which eventually leads to very large weights. This instability can be suppressed by weight normalization, at the cost of bias in the learning rule. We confirm numerically that a similar instability, and to a lesser extent scalability, exist in deep nonlinear networks trained on both a motor control task and image classification tasks. Our study highlights the limitations and potential of node perturbation as a biologically plausible learning rule in the brain.

## [Near-Isometric Properties of Kronecker-Structured Random Tensor Embeddings](#)

- Qijia Jiang
- abstract@[open-review](#): We give uniform concentration inequality for random tensor acting on rank-1 Kronecker structured signals, which parallels a Gordon-type inequality for this class of tensor structured data. Two variants of the random embedding are considered, where the embedding dimension depends on explicit quantities characterizing the complexity of the signal. To appreciate the tools developed herein, we illustrate with two applications from signal recovery and optimization.

## [MACE: Higher Order Equivariant Message Passing Neural Networks for Fast and Accurate Force Fields](#)

- Ilyes Batatia · David P Kovacs · Gregor Simm · Christoph Ortner · Gabor Csanyi
- abstract@[open-review](#): Creating fast and accurate force fields is a long-standing challenge in computational chemistry and materials science. Recently, Equivariant Message Passing Neural Networks (MPNNs) have emerged as a powerful tool for building machine learning interatomic potentials, outperforming other approaches in terms of accuracy. However, they suffer from high computational cost and poor scalability. Moreover, most MPNNs only pass two-body messages leading to an intricate relationship between the number of layers and the expressivity of the features. This work introduces MACE, a new equivariant MPNN model that uses higher order messages, and demonstrates that this leads to an improved learning law. We show that by using four-body messages, the required number of message passing iterations reduces to just one, resulting in a fast and highly parallelizable model, reaching or exceeding state of the art accuracy on the rMD17 and 3BPA benchmark tasks.

## [Causal Imitation Learning with Unobserved Contexts](#)

- Gokul Swamy · Sanjiban Choudhury · J. Bagnell · Steven Wu
- abstract@[open-review](#): We consider imitation learning problems where the expert has access to a per-episode context that is hidden from the learner, both in the demonstrations and at test-time. While the learner might not be able to accurately reproduce expert behavior early on in an episode, by considering the entire history of states and actions, they might be able to eventually identify the context and act as the expert would. We show that on-policy imitation learning algorithms (with or without access to a queryable expert) are better equipped to handle these sorts of asymptotically realizable problems than off-policy methods and are able to avoid the latching behavior that plagues the latter. We conduct experiments in a toy bandit domain that show that there

exist sharp phase transitions of whether off-policy approaches are able to match expert performance asymptotically, in contrast to the uniformly good performance of on-policy approaches. We demonstrate that on several continuous control tasks, on-policy approaches are able to use history to identify the context while off-policy approaches are unable to do so.

## [Learning Two-Player Markov Games: Neural Function Approximation and Correlated Equilibrium](#)

- Chris Junchi Li · Dongruo Zhou · Quanquan Gu · Michael Jordan
- abstract@[open-review](#): We consider learning Nash equilibria in two-player zero-sum Markov Games with nonlinear function approximation, where the action-value function is approximated by a function in a Reproducing Kernel Hilbert Space (RKHS). The key challenge is how to do exploration in the high-dimensional function space. We propose a novel online learning algorithm to find a Nash equilibrium by minimizing the duality gap. At the core of our algorithms are upper and lower confidence bounds that are derived based on the principle of optimism in the face of uncertainty. We prove that our algorithm is able to attain an  $\mathcal{O}(\sqrt{T})$  regret with polynomial computational complexity, under very mild assumptions on the reward function and the underlying dynamic of the Markov Games. We also propose several extensions of our algorithm, including an algorithm with Bernstein-type bonus that can achieve a tighter regret bound, and another algorithm for model misspecification that can be applied to neural network function approximation.

## [Probable Domain Generalization via Quantile Risk Minimization](#)

- Cian Eastwood · Alexander Robey · Shashank Singh · Julius von Kägelgen · Hamed Hassani · George J. Pappas · Bernhard Schölkopf
- abstract@[open-review](#): Domain generalization (DG) leverages labeled training data from multiple domains with the goal of generalizing to related test domains. To achieve this, DG is commonly formulated as a worst-case optimization problem over the set of all possible domains. However, this worst-case problem is generally intractable and, with adversarial shifts extremely unlikely in practice, leads to overly-conservative solutions. In fact, a recent study found that no DG algorithm outperformed empirical risk minimization in terms of average performance over test domains. To address these shortcomings, we propose a probabilistic framework for DG, which we call Probable Domain Generalization, and advocate for predictors that perform well with high probability rather than in the worst-case or on-average. Our key idea is that distribution shifts seen during training should inform us of probable shifts at test time. To achieve this, we explicitly relate training and test domains as draws from the same underlying meta-distribution, and propose a new optimization problem---Quantile Risk Minimization (QRM)---which requires that predictors generalize with high probability. We then prove that, given sufficiently many domains and samples, the empirical version (EQRM) produces predictors that generalize to new domains with the desired probability. We also show that EQRM recovers the causal predictor as the desired probability of generalization approaches one. In our experiments, we introduce a new evaluation protocol for DG, which underscores the importance of multiple test domains for evaluating the quantile performance of DG algorithms, and we show that our algorithms outperform strong DG baselines on real and synthetic data.

## [Teacher Forcing Recovers Reward Functions for Text Generation](#)

- Yongchang Hao · Yuxin Liu · Lili Mou
- abstract@[open-review](#): Through the lens of inverse reinforcement learning, we propose a method to derive the reward function from models trained with the teacher-forcing objective. The method gives us a step-wise reward function for sequence generation and enables reinforcement learning for text generation. Compared to previous approaches, our method is task-agnostic and does not require any human heuristics. In addition, we develop a stable training strategy based on the policy gradient method, which leads to further improvement on non-parallel datasets when combined with our proposed reward. The empirical results show that our method outperforms several previous works, which confirms the effectiveness of our reward function. The code is publicly available.

## [Defining and Characterizing Reward Gaming](#)

- Joar Skalse · Nikolaus Howe · Dmitrii Krasheninnikov · David Krueger
- abstract@[open-review](#): We provide the first formal definition of \textbf{reward gaming}, a phenomenon where optimizing an imperfect \emph{proxy reward function},  $\tilde{R}$ , leads to poor performance according to an intended reward function,  $R$ . We say that a proxy is \emph{ungameable} if increasing the expected proxy return can never decrease the expected intended return. Intuitively, it should be possible to create an ungameable proxy by overlooking fine-grained distinctions between roughly equivalent outcomes, but we show this is usually not the case. A key insight is that the linearity of reward (as a function of state-action visit counts) makes ungameability a very strong condition. In particular, for the set of all stochastic policies, two reward functions can only be ungameable if one of them is constant. We thus turn our attention to deterministic policies and finite sets of stochastic policies, and establish necessary and sufficient conditions for the existence of non-trivial ungameable pairs of reward functions. Our results reveal a tension between using reward functions to specify narrow tasks and aligning AI systems with human values. We provide the first formal definition of \textbf{reward gaming}, a phenomenon where optimizing an imperfect \emph{proxy reward function},  $\tilde{R}$ , leads to poor performance according to a true reward function,  $R$ . We say that a proxy is \emph{ungameable} if increasing the expected proxy return can never decrease the expected true return. Intuitively, it should be possible to create an ungameable proxy by overlooking fine-grained distinctions between roughly equivalent outcomes, but we show this is usually not the case. A key insight is that the linearity of reward (as a function of state-action visit counts) makes ungameability a very strong condition. In particular, for the set of all stochastic policies, two reward functions can only be ungameable if one of them is constant. We thus turn our attention to deterministic policies and finite sets of stochastic policies, where non-trivial ungameable pairs always exist, and establish necessary and sufficient conditions for the existence of simplifications, an important special case of ungameability. Our results reveal a tension between using reward functions to specify narrow tasks and aligning AI systems with human values.

## [Embed and Emulate: Learning to estimate parameters of dynamical systems with uncertainty quantification](#)

- Ruoxi Jiang · Rebecca Willett
- abstract@[open-review](#): This paper explores learning emulators for parameter estimation with uncertainty estimation of high-dimensional dynamical systems. We assume access to a computationally complex simulator that inputs a candidate parameter and outputs a corresponding multi-channel time series. Our task is to accurately estimate a range of likely values of the underlying parameters. Standard iterative approaches necessitate running the simulator many times, which is computationally prohibitive. This paper describes a novel framework for learning feature embeddings of observed dynamics jointly with an emulator that can replace high-cost simulators. Leveraging a contrastive learning approach, our method exploits intrinsic data properties within and across parameter and trajectory domains. On a coupled 396-dimensional multiscale Lorenz 96 system, our method significantly outperforms a typical parameter estimation method based on predefined metrics and a classical numerical simulator, and with only 3.5% of the baseline's computation time. Ablation studies highlight the potential of explicitly designing learned emulators for parameter estimation by leveraging contrastive learning.

## [DGD<sup>2</sup>: A Linearly Convergent Distributed Algorithm For High-dimensional Statistical Recovery](#)

- Marie Maros · Gesualdo Scutari
- abstract@[open-review](#): We study linear regression from data distributed over a network of agents (with no master node) under high-dimensional scaling, which allows the ambient dimension to grow faster than the sample size. We propose a novel decentralization of the projected gradient algorithm whereby agents iteratively update their local estimates by a double-mixing mechanism, which suitably combines averages of iterates and gradients of neighbouring nodes. Under standard assumptions on the statistical model and network connectivity, the proposed method enjoys global linear convergence up to the statistical precision of the model. This improves on guarantees of (plain) DGD algorithms, whose iteration complexity grows undesirably with

the ambient dimension. Our technical contribution is a novel convergence analysis that resembles (albeit different) algorithmic stability arguments extended to high-dimensions and distributed setting, which is of independent interest.

## [Draft-and-Revise: Effective Image Generation with Contextual RQ-Transformer](#)

- Doyup Lee · Chiheon Kim · Saehoon Kim · Minsu Cho · WOOK SHIN HAN
- abstract@[open-review](#): Although autoregressive models have achieved promising results on image generation, their unidirectional generation process prevents the resultant images from fully reflecting global contexts. To address the issue, we propose an effective image generation framework of \emph{Draft-and-Revise} with \emph{Contextual RQ-transformer} to consider global contexts during the generation process. As a generalized VQ-VAE, RQ-VAE first represents a high-resolution image as a sequence of discrete code stacks. After code stacks in the sequence are randomly masked, Contextual RQ-Transformer is trained to infill the masked code stacks based on the unmasked contexts of the image. Then, we propose the two-phase decoding, Draft-and-Revise, for Contextual RQ-Transformer to generates an image, while fully exploiting the global contexts of the image during the generation process. Specifically, in the \emph{draft} phase, our model first focuses on generating diverse images despite rather low quality. Then, in the \emph{revise} phase, the model iteratively improves the quality of images, while preserving the global contexts of generated images. In experiments, our method achieves state-of-the-art results on conditional image generation. We also validate that the Draft-and-Revise decoding can achieve high performance by effectively controlling the quality-diversity trade-off in image generation.

## [Variational Model Perturbation for Source-Free Domain Adaptation](#)

- Mengmeng Jing · Xiantong Zhen · Jingjing Li · Cees Snoek
- abstract@[open-review](#): We aim for source-free domain adaptation, where the task is to deploy a model pre-trained on source domains to target domains. The challenges stem from the distribution shift from the source to the target domain, coupled with the unavailability of any source data and labeled target data for optimization. Rather than fine-tuning the model by updating the parameters, we propose to perturb the source model to achieve adaptation to target domains. We introduce perturbations into the model parameters by variational Bayesian inference in a probabilistic framework. By doing so, we can effectively adapt the model to the target domain while largely preserving the discriminative ability. Importantly, we demonstrate the theoretical connection to learning Bayesian neural networks, which proves the generalizability of the perturbed model to target domains. To enable more efficient optimization, we further employ a parameter sharing strategy, which substantially reduces the learnable parameters compared to a fully Bayesian neural network. Our model perturbation provides a new probabilistic way for domain adaptation which enables efficient adaptation to target domains while maximally preserving knowledge in source models. Experiments on several source-free benchmarks under three different evaluation settings verify the effectiveness of the proposed variational model perturbation for source-free domain adaptation.

## [Optimal Rates for Regularized Conditional Mean Embedding Learning](#)

- Zhu Li · Dimitri Meunier · Arthur Gretton
- abstract@[open-review](#): We address the consistency of a kernel ridge regression estimate of the conditional mean embedding (CME), which is an embedding of the conditional distribution of  $\mathbf{Y}$  given  $\mathbf{X}$  into a target reproducing kernel Hilbert space  $\mathcal{H}_Y$ . The CME allows us to take conditional expectations of target RKHS functions, and has been employed in nonparametric causal and Bayesian inference. We address the misspecified setting, where the target CME is in the space of Hilbert-Schmidt operators acting from an input interpolation space between  $\mathcal{H}_X$  and  $L_2$ , to  $\mathcal{H}_Y$ . This space of operators is shown to be isomorphic to a newly defined vector-valued interpolation space. Using this isomorphism, we derive a novel and adaptive statistical learning rate for the empirical CME estimator under the misspecified setting. Our analysis reveals that our rates match the optimal  $O(\log n / n)$  rates without assuming  $\mathcal{H}_Y$  to be finite dimensional. We further establish a lower bound on the learning rate, which shows that the obtained upper bound is optimal.

## [On the SDEs and Scaling Rules for Adaptive Gradient Algorithms](#)

- Sadhika Malladi · Kaifeng Lyu · Abhishek Panigrahi · Sanjeev Arora
- abstract@[open-review](#): Approximating Stochastic Gradient Descent (SGD) as a Stochastic Differential Equation (SDE) has allowed researchers to enjoy the benefits of studying a continuous optimization trajectory while carefully preserving the stochasticity of SGD. Analogous study of adaptive gradient methods, such as RMSprop and Adam, has been challenging because there were no rigorously proven SDE approximations for these methods. This paper derives the SDE approximations for RMSprop and Adam, giving theoretical guarantees of their correctness as well as experimental validation of their applicability to common large-scaling vision and language settings. A key practical result is the derivation of a square root scaling rule to adjust the optimization hyperparameters of RMSprop and Adam when changing batch size, and its empirical validation in deep learning settings.

## [A Theoretical Understanding of Gradient Bias in Meta-Reinforcement Learning](#)

- Bo Liu · Xidong Feng · Jie Ren · Luo Mai · Rui Zhu · Haifeng Zhang · Jun Wang · Yaodong Yang
- abstract@[open-review](#): Gradient-based Meta-RL (GMRL) refers to methods that maintain two-level optimisation procedures wherein the outer-loop meta-learner guides the inner-loop gradient-based reinforcement learner to achieve fast adaptations. In this paper, we develop a unified framework that describes variations of GMRL algorithms and points out that existing stochastic meta-gradient estimators adopted by GMRL are actually \textbf{biased}. Such meta-gradient bias comes from two sources: 1) the compositional bias incurred by the two-level problem structure, which has an upper bound of  $\mathcal{O}(K\alpha^K)\hat{\sigma}\sqrt{In}\tau^{0.5}$  inner-loop update step  $K$ , learning rate  $\alpha$ , estimate variance  $\hat{\sigma}^2$  and sample size  $n$ , and 2) the multi-step Hessian estimation bias  $\hat{\Delta}_H$  due to the use of autodiff, which has a polynomial impact  $\mathcal{O}(K(K-1)\hat{\Delta}_H^{K-1})$  on the meta-gradient bias. We study tabular MDPs empirically and offer quantitative evidence that testifies our theoretical findings on existing stochastic meta-gradient estimators. Furthermore, we conduct experiments on Iterated Prisoner's Dilemma and Atari games to show how other methods such as off-policy learning and low-bias estimator can help fix the gradient bias for GMRL algorithms in general.

## [Wavelet Feature Maps Compression for Image-to-Image CNNs](#)

- Shahaf E. Finder · Yair Zohav · Maor Ashkenazi · Eran Treister
- abstract@[open-review](#): Convolutional Neural Networks (CNNs) are known for requiring extensive computational resources, and quantization is among the best and most common methods for compressing them. While aggressive quantization (i.e., less than 4-bits) performs well for classification, it may cause severe performance degradation in image-to-image tasks such as semantic segmentation and depth estimation. In this paper, we propose Wavelet Compressed Convolution (WCC)---a novel approach for high-resolution activation maps compression integrated with point-wise convolutions, which are the main computational cost of modern architectures. To this end, we use an efficient and hardware-friendly Haar-wavelet transform, known for its effectiveness in image compression, and define the convolution on the compressed activation map. We experiment on various tasks, that benefit from high-resolution input, and by combining WCC with light quantization, we achieve compression rates equivalent to 1-4bit activation quantization with relatively small and much more graceful degradation in performance.

## [Paraphrasing Is All You Need for Novel Object Captioning](#)

- Cheng-Fu Yang · Yao-Hung Hubert Tsai · Wan-Cyuan Fan · Russ Salakhutdinov · Louis-Philippe Morency · Frank Wang
- abstract@[open-review](#): Novel object captioning (NOC) aims to describe images containing objects without observing their ground truth captions during training. Due to the absence of caption annotation, captioning models cannot be directly optimized via sequence-to-sequence training or CIDEr optimization. As a result, we present Paraphrasing-to-Captioning (P2C), a two-stage learning framework for NOC, which would heuristically optimize the output captions via paraphrasing. With P2C, the captioning model first learns paraphrasing from a language model pre-trained on text-only corpus, allowing expansion of the word bank for improving linguistic fluency. To further enforce the output caption sufficiently describing the visual content of the input image, we perform self-paraphrasing for the captioning model with fidelity and adequacy objectives introduced. Since no ground truth captions are available for novel object images during training, our P2C leverages cross-modality (image-text) association modules to ensure the above caption characteristics can be properly preserved. In the experiments, we not only show that our P2C achieves state-of-the-art performances on nocaps and COCO Caption datasets, we also verify the effectiveness and flexibility of our learning framework by replacing language and cross-modality association models for NOC. Implementation details and code are available in the supplementary materials.

## [Redundancy-Free Message Passing for Graph Neural Networks](#)

- Rongqin Chen · Shenghui Zhang · Leong Hou U · Ye Li
- abstract@[open-review](#): Graph Neural Networks (GNNs) resemble the Weisfeiler-Lehman (1-WL) test, which iteratively update the representation of each node by aggregating information from WL-tree. However, despite the computational superiority of the iterative aggregation scheme, it introduces redundant message flows to encode nodes. We found that the redundancy in message passing prevented conventional GNNs from propagating the information of long-length paths and learning graph similarities. In order to address this issue, we proposed Redundancy-Free Graph Neural Network (RFGNN), in which the information of each path (of limited length) in the original graph is propagated along a single message flow. Our rigorous theoretical analysis demonstrates the following advantages of RFGNN: (1) RFGNN is strictly more powerful than 1-WL; (2) RFGNN efficiently propagate structural information in original graphs, avoiding the over-squashing issue; and (3) RFGNN could capture subgraphs at multiple levels of granularity, and are more likely to encode graphs with closer graph edit distances into more similar representations. The experimental evaluation of graph-level prediction benchmarks confirmed our theoretical assertions, and the performance of the RFGNN can achieve the best results in most datasets.

## [Insights into Pre-training via Simpler Synthetic Tasks](#)

- Yuhuai Wu · Felix Li · Percy Liang
- abstract@[open-review](#): Pre-training produces representations that are effective for a wide range of down-stream tasks, but it is still unclear what properties of pre-training are necessary for effective gains. Notably, recent work shows that even pre-training on synthetic tasks can achieve significant gains in downstream tasks. In this work, we perform three experiments that iteratively simplify pre-training and show that the simplifications still retain much of its gains. First, building on prior work, we perform a systematic evaluation of three existing synthetic pre-training methods on six downstream tasks. We find the best synthetic pre-training method, LIME, attains an average of 67% of the benefits of natural pre-training. Second, to our surprise, we find that pre-training on a simple and generic synthetic task defined by the Set function achieves 65% of the benefits, almost matching LIME. Third, we find that 39% of the benefits can be attained by using merely the parameter statistics of synthetic pre-training.

## [An Information-Theoretic Framework for Deep Learning](#)

- Hong Jun Jeon · Benjamin Van Roy
- abstract@[open-review](#): Each year, deep learning demonstrate new and improved empirical results with deeper and wider neural networks. Meanwhile, with existing theoretical frameworks, it is difficult to analyze networks deeper than two layers without resorting to counting parameters or encountering sample complexity bounds that are exponential in depth. Perhaps it may be fruitful to try to analyze modern machine learning under a different lens. In this paper, we propose a novel information-theoretic framework with its own notions of regret and sample complexity for analyzing the data requirements of machine learning. We use this framework to study the sample complexity of learning from data generated by deep ReLU neural networks and deep networks that are infinitely wide but have a bounded sum of weights. We establish that the sample complexity of learning under these data generating processes is at most linear and quadratic, respectively, in network depth.

## [Rate-Optimal Online Convex Optimization in Adaptive Linear Control](#)

- Asaf Benjamin Cassel · Alon Peled-Cohen · Tomer Koren
- abstract@[open-review](#): We consider the problem of controlling an unknown linear dynamical system under adversarially-changing convex costs and full feedback of both the state and cost function. We present the first computationally-efficient algorithm that attains an optimal  $\sqrt{T}$ -regret rate compared to the best stabilizing linear controller in hindsight, while avoiding stringent assumptions on the costs such as strong convexity. Our approach is based on a careful design of non-convex lower confidence bounds for the online costs, and uses a novel technique for computationally-efficient regret minimization of these bounds that leverages their particular non-convex structure.

## [AVLEN: Audio-Visual-Language Embodied Navigation in 3D Environments](#)

- Sudipta Paul · Amit Roy-Chowdhury · Anoop Cherian
- abstract@[open-review](#): Recent years have seen embodied visual navigation advance in two distinct directions: (i) in equipping the AI agent to follow natural language instructions, and (ii) in making the navigable world multimodal, e.g., audio-visual navigation. However, the real world is not only multimodal, but also often complex, and thus in spite of these advances, agents still need to understand the uncertainty in their actions and seek instructions to navigate. To this end, we present AVLEN -- an interactive agent for Audio-Visual-Language Embodied Navigation. Similar to audio-visual navigation tasks, the goal of our embodied agent is to localize an audio event via navigating the 3D visual world; however, the agent may also seek help from a human (oracle), where the assistance is provided in free-form natural language. To realize these abilities, AVLEN uses a multimodal hierarchical reinforcement learning backbone that learns: (a) high-level policies to choose either audio-cues for navigation or to query the oracle, and (b) lower-level policies to select navigation actions based on its audio-visual and language inputs. The policies are trained via rewarding for the success on the navigation task while minimizing the number of queries to the oracle. To empirically evaluate AVLEN, we present experiments on the SoundSpaces framework for semantic audio-visual navigation tasks. Our results show that equipping the agent to ask for help leads to a clear improvement in performances, especially in challenging cases, e.g., when the sound is unheard during training or in the presence of distractor sounds.

## [A Consistent, Scalable, and Differentiable L<sub>p</sub> Canonical Calibration Error Estimator](#)

- Teodora Popordanoska · Raphael Sayer · Matthew Blaschko
- abstract@[open-review](#): Calibrated probabilistic classifiers are models whose predicted probabilities can directly be interpreted as uncertainty estimates. It has been shown recently that deep neural networks are poorly calibrated and tend to output overconfident predictions. As a remedy, we propose a low-bias, trainable calibration error estimator based on Dirichlet kernel density estimates, which asymptotically converges to the true  $L_p$  calibration error. This novel estimator enables us to achieve the strongest notion of multiclass calibration, called canonical calibration, while other common calibration methods are tractable only for top-label and marginal calibration. The computational complexity of our estimator is  $O(n^2)$ , while the convergence rate is  $O(n^{-1/2})$ , and it is unbiased up to  $O(n^{-2})$  achieved by a geometric series debiasing scheme. In practice, this means that the estimator can be applied to small subsets of the data, enabling efficient estimation and mini-batch updates. The proposed method has a natural choice of kernel, and can be used to generate consistent estimates of other quantities based on conditional expectation, such as the

sharpness of a probabilistic classifier. Empirical results validate the correctness of our estimator, and demonstrate its utility in canonical calibration error estimation and calibration error regularized risk minimization.

## [The Burer-Monteiro SDP method can fail even above the Barvinok-Pataki bound](#)

- Liam O'Carroll · Vaidehi Srinivas · Aravindan Vijayaraghavan
- abstract@[open-review](#): The most widely used technique for solving large-scale semidefinite programs (SDPs) in practice is the non-convex Burer-Monteiro method, which explicitly maintains a low-rank SDP solution for memory efficiency. There has been much recent interest in obtaining a better theoretical understanding of the Burer-Monteiro method. When the rank  $r$  of the SDP solution is above the Barvinok-Pataki bound (where a globally optimal solution of rank at most  $(r)$  is guaranteed to exist), a recent line of work established convergence to a global optimum for generic or smoothed instances of the problem. However, it was open whether there even exists an instance in this regime where the Burer-Monteiro method fails. We prove that the Burer-Monteiro method can fail for the Max-Cut SDP on  $n$  vertices when the rank is above the Barvinok-Pataki bound ( $r \geq \sqrt{2n}$ ). We provide a family of instances that have spurious local minima even when the rank  $r = n/2$ . Combined with existing guarantees, this settles the question of the existence of spurious local minima for the Max-Cut formulation in all ranges of the rank, and justifies the use of beyond-worst-case paradigms like smoothed analysis to obtain guarantees for the Burer-Monteiro method.

## [Learning in Congestion Games with Bandit Feedback](#)

- Qiwen Cui · Zhihan Xiong · Maryam Fazel · Simon Du
- abstract@[open-review](#): Learning the Nash equilibrium is a central problem in multi-agent systems. In this paper, we investigate congestion games, a class of games with benign theoretical structure and broad real-world applications. We first propose a centralized algorithm based on the optimism in the face of uncertainty principle for congestion games with (semi-)bandit feedback, and obtain finite-sample guarantees. Then we propose a decentralized algorithm via a novel combination of the Frank-Wolfe method and G-optimal design. By exploiting the structure of the congestion game, we show the sample complexity of both algorithms depends only polynomially on the number of players and the number of facilities, but not the size of the action set, which can be exponentially large in terms of the number of facilities. We further define a new problem class, Markov congestion games, which allows us to model the non-stationarity in congestion games. We propose a centralized algorithm for Markov congestion games, whose sample complexity again has only polynomial dependence on all relevant problem parameters, but not the action size.

## [Revisiting Non-Parametric Matching Cost Volumes for Robust and Generalizable Stereo Matching](#)

- Kelvin Cheng · Tianfu Wu · Zhebin Zhang · Hongyu Sun · Christopher Healey
- abstract@[open-review](#): Stereo matching is a classic challenging problem in computer vision, which has recently witnessed remarkable progress by Deep Neural Networks (DNNs). This paradigm shift leads to two interesting questions that have not been addressed well. First, it is unclear whether stereo matching DNNs really learn to perform matching well. This paper studies this problem from the lens of adversarial attacks. It presents a method of learning stereo-constrained photometrically-consistent attacks, which by design are weaker adversarial attacks. State-of-the-art stereo matching DNNs are, however, vulnerable against them. This observation suggests that DNNs may not learn to perform matching well in the sense that they should otherwise achieve potentially even better after stereo-constrained perturbations are introduced. Second, stereo matching DNNs are typically trained under the simulation-to-real (Sim2Real) pipeline due to the data hungry of DNNs. Thus, alleviating the impacts of the Sim2Real photometric gap in stereo matching DNNs becomes a pressing need. Towards adversarially robust and domain generalizable stereo matching, this paper proposes to rethink the role of DNNs. It presents a method that casts stereo matching as a cost aggregation problem (solved by training a DNN) over a non-parametric cost volume (that truly focuses on matching) with parametric contextual features. In experiments, the proposed method is tested in the SceneFlow dataset, the KITTI2015 dataset, and the Middlebury dataset. It significantly improves the adversarial robustness, while retaining accuracy performance comparable to state-of-the-art methods. It also shows better Sim2Real generalizability.

## [Lifelong Neural Predictive Coding: Learning Cumulatively Online without Forgetting](#)

- Alex Ororbia · Ankur Mali · C Lee Giles · Daniel Kifer
- abstract@[open-review](#): In lifelong learning systems based on artificial neural networks, one of the biggest obstacles is the inability to retain old knowledge as new information is encountered. This phenomenon is known as catastrophic forgetting. In this paper, we propose a new kind of connectionist architecture, the Sequential Neural Coding Network, that is robust to forgetting when learning from streams of data points and, unlike networks of today, does not learn via the popular back-propagation of errors. Grounded in the neurocognitive theory of predictive processing, our model adapts synapses in a biologically-plausible fashion while another neural system learns to direct and control this cortex-like structure by mimicking some of task-executive control functionality of the basal ganglia. In our experiments, we demonstrate that our self-organizing system experiences significantly less forgetting compared to standard neural models, outperforming a swath of previously proposed methods, including rehearsal/data buffer-based methods, on both standard (SplitMNIST, Split Fashion MNIST, etc.) and custom benchmarks even though it is trained in a stream-like fashion. Our work offers evidence that emulating mechanisms in real neuronal systems, e.g., local learning, lateral competition, can yield new directions and possibilities for tackling the grand challenge of lifelong machine learning.

## [Generalized One-shot Domain Adaption of Generative Adversarial Networks](#)

- Zicheng Zhang · Yinglu Liu · Congying Han · Tiande Guo · Ting Yao · Tao Mei
- abstract@[open-review](#): The adaption of Generative Adversarial Network (GAN) aims to transfer a pre-trained GAN to a given domain with limited training data. In this paper, we focus on the one-shot case, which is more challenging and rarely explored in previous works. We consider that the adaptation from source domain to target domain can be decoupled into two parts: the transfer of global style like texture and color, and the emergence of new entities that do not belong to the source domain. While previous works mainly focus on the style transfer, we propose a novel and concise framework to address the generalized one-shot adaption task for both style and entity transfer, in which a reference image and its binary entity mask are provided. Our core objective is to constrain the gap between the internal distributions of the reference and syntheses by sliced Wasserstein distance. To better achieve it, style fixation is used at first to roughly obtain the exemplary style, and an auxiliary network is introduced to the original generator to disentangle entity and style transfer. Besides, to realize cross-domain correspondence, we propose the variational Laplacian regularization to constrain the smoothness of the adapted generator. Both quantitative and qualitative experiments demonstrate the effectiveness of our method in various scenarios.

## [Stochastic Online Learning with Feedback Graphs: Finite-Time and Asymptotic Optimality](#)

- Teodor Vanislavov Marinov · Mehryar Mohri · Julian Zimmert
- abstract@[open-review](#): We revisit the problem of stochastic online learning with feedback graphs, with the goal of devising algorithms that are optimal, up to constants, both asymptotically and in finite time. We show that, surprisingly, the notion of optimal finite-time regret is not a uniquely defined property in this context and that, in general, it is decoupled from the asymptotic rate. We discuss alternative choices and propose a notion of finite-time optimality that we argue is meaningful. For that notion, we give an algorithm that admits quasi-optimal regret both in finite-time and asymptotically.

## [Distribution-Informed Neural Networks for Domain Adaptation Regression](#)

- Jun Wu · Jingrui He · Sheng Wang · Kaiyu Guan · Elizabeth Ainsworth
- abstract@[open-review](#): In this paper, we study the problem of domain adaptation regression, which learns a regressor for a target domain by leveraging the knowledge from a relevant source domain. We start by proposing a distribution-informed neural network, which aims to build distribution-aware relationship of inputs and outputs from different domains. This allows us to develop a simple domain adaptation regression framework, which subsumes popular domain adaptation approaches based on domain invariant representation learning, reweighting, and adaptive Gaussian process. The resulting findings not only explain the connections of existing domain adaptation approaches, but also motivate the efficient training of domain adaptation approaches with overparameterized neural networks. We also analyze the convergence and generalization error bound of our framework based on the distribution-informed neural network. Specifically, our generalization bound focuses explicitly on the maximum mean discrepancy in the RKHS induced by the neural tangent kernel of distribution-informed neural network. This is in sharp contrast to the existing work which relies on domain discrepancy in the latent feature space heuristically formed by one or several hidden neural layers. The efficacy of our framework is also empirically verified on a variety of domain adaptation regression benchmarks.

## [Rare Gems: Finding Lottery Tickets at Initialization](#)

- Kartik Sreenivasan · Jy-yong Sohn · Liu Yang · Matthew Grinde · Alliot Nagle · Hongyi Wang · Eric Xing · Kangwook Lee · Dimitris Papailiopoulos
- abstract@[open-review](#): Large neural networks can be pruned to a small fraction of their original size, with little loss in accuracy, by following a time-consuming "train, prune, re-train" approach. Frankle & Carbin conjecture that we can avoid this by training lottery tickets, i.e., special sparse subnetworks found at initialization, that can be trained to high accuracy. However, a subsequent line of work presents concrete evidence that current algorithms for finding trainable networks at initialization, fail simple baseline comparisons, e.g., against training random sparse subnetworks. Finding lottery tickets that train to better accuracy compared to simple baselines remains an open problem. In this work, we resolve this open problem by proposing Gem-Miner which finds lottery tickets at initialization that beat current baselines. Gem-Miner finds lottery tickets trainable to accuracy competitive or better than Iterative Magnitude Pruning (IMP), and does so up to \$19\times\$ faster.

## [Optimal Transport of Classifiers to Fairness](#)

- Maarten Buyl · Tijl De Bie
- abstract@[open-review](#): In past work on fairness in machine learning, the focus has been on forcing the prediction of classifiers to have similar statistical properties for people of different demographics. To reduce the violation of these properties, fairness methods usually simply rescale the classifier scores, ignoring similarities and dissimilarities between members of different groups. Yet, we hypothesize that such information is relevant in quantifying the unfairness of a given classifier. To validate this hypothesis, we introduce Optimal Transport to Fairness (OTF), a method that quantifies the violation of fairness constraints as the smallest Optimal Transport cost between a probabilistic classifier and any score function that satisfies these constraints. For a flexible class of linear fairness constraints, we construct a practical way to compute OTF as a differentiable fairness regularizer that can be added to any standard classification setting. Experiments show that OTF can be used to achieve an improved trade-off between predictive power and fairness.

## [AutoML Two-Sample Test](#)

- Jonas Käbler · Vincent Stimper · Simon Buchholz · Krikamol Muandet · Bernhard Schölkopf
- abstract@[open-review](#): Two-sample tests are important in statistics and machine learning, both as tools for scientific discovery as well as to detect distribution shifts. This led to the development of many sophisticated test procedures going beyond the standard supervised learning frameworks, whose usage can require specialized knowledge about two-sample testing. We use a simple test that takes the mean discrepancy of a witness function as the test statistic and prove that minimizing a squared loss leads to a witness with optimal testing power. This allows us to leverage recent advancements in AutoML. Without any user input about the problems at hand, and using the same method for all our experiments, our AutoML two-sample test achieves competitive performance on a diverse distribution shift benchmark as well as on challenging two-sample testing problems.

## [Perfect Sampling from Pairwise Comparisons](#)

- Dimitris Fotakis · Alkis Kalavasis · Christos Tzamos
- abstract@[open-review](#): In this work, we study how to efficiently obtain perfect samples from a discrete distribution  $\mathcal{D}$  given access only to pairwise comparisons of elements of its support. Specifically, we assume access to samples  $(x, S)$ , where  $S$  is drawn from a distribution over sets  $\mathcal{Q}$  (indicating the elements being compared), and  $x$  is drawn from the conditional distribution  $\mathcal{D}_S$  (indicating the winner of the comparison) and aim to output a clean sample  $y$  distributed according to  $\mathcal{D}$ . We mainly focus on the case of pairwise comparisons where all sets  $S$  have size 2. We design a Markov chain whose stationary distribution coincides with  $\mathcal{D}$  and give an algorithm to obtain exact samples using the technique of Coupling from the Past. However, the sample complexity of this algorithm depends on the structure of the distribution  $\mathcal{D}$  and can be even exponential in the support of  $\mathcal{D}$  in many natural scenarios. Our main contribution is to provide an efficient exact sampling algorithm whose complexity does not depend on the structure of  $\mathcal{D}$ . To this end, we give a parametric Markov chain that mixes significantly faster given a good approximation to the stationary distribution. We can obtain such an approximation using an efficient learning from pairwise comparisons algorithm (Shah et al., JMLR 17, 2016). Our technique for speeding up sampling from a Markov chain whose stationary distribution is approximately known is simple, general and possibly of independent interest.

## [Listen to Interpret: Post-hoc Interpretability for Audio Networks with NMF](#)

- Jayneel Parekh · Sanjeel Parekh · Pavlo Mozharovskyi · Florence d'Alchău-Buc · Gaël Richard
- abstract@[open-review](#): This paper tackles post-hoc interpretability for audio processing networks. Our goal is to interpret decisions of a trained network in terms of high-level audio objects that are also listenable for the end-user. To this end, we propose a novel interpreter design that incorporates non-negative matrix factorization (NMF). In particular, a regularized interpreter module is trained to take hidden layer representations of the targeted network as input and produce time activations of pre-learnt NMF components as intermediate outputs. Our methodology allows us to generate intuitive audio-based interpretations that explicitly enhance parts of the input signal most relevant for a network's decision. We demonstrate our method's applicability on popular benchmarks, including a real-world multi-label classification task.

## [Temporally-Consistent Survival Analysis](#)

- Lucas Maystre · Daniel Russo
- abstract@[open-review](#): We study survival analysis in the dynamic setting: We seek to model the time to an event of interest given sequences of states. Taking inspiration from temporal-difference learning, a central idea in reinforcement learning, we develop algorithms that estimate a discrete-time survival model by exploiting a temporal-consistency condition. Intuitively, this condition captures the fact that the survival distribution at consecutive states should be similar, accounting for the delay between states. Our method can be combined with any parametric survival model and naturally accommodates right-censored observations. We demonstrate empirically that it achieves better sample-efficiency and predictive performance compared to approaches that directly regress the observed survival outcome.

## [Efficient Scheduling of Data Augmentation for Deep Reinforcement Learning](#)

- Byungchan Ko · Jungseul Ok
- abstract@[open-review](#): In deep reinforcement learning (RL), data augmentation is widely considered as a tool to induce a set of useful priors about semantic consistency and improve sample efficiency and generalization performance. However, even when the prior is useful for generalization, distilling it to RL agent often interferes with RL training and degenerates sample efficiency. Meanwhile, the agent is forgetful of the prior due to the non-stationary nature of RL. These observations suggest two extreme schedules of distillation: (i) over the entire training; or (ii) only at the end. Hence, we devise a stand-alone network distillation method to inject the consistency prior at any time (even after RL), and a simple yet efficient framework to automatically schedule the distillation. Specifically, the proposed framework first focuses on mastering train environments regardless of generalization by adaptively deciding which {vit or no} augmentation to be used for the training. After this, we add the distillation to extract the remaining benefits for generalization from all the augmentations, which requires no additional new samples. In our experiments, we demonstrate the utility of the proposed framework, in particular, that considers postponing the augmentation to the end of RL training.

## [Black-Box Generalization](#)

- Konstantinos Nikolakakis · Farzin Haddadpour · Dionysis Kalogerias · Amin Karbasi
- abstract@[open-review](#): We provide the first generalization error analysis for black-box learning through derivative-free optimization. Under the assumption of a Lipschitz and smooth unknown loss, we consider the Zeroth-order Stochastic Search (ZoSS) algorithm, that updates a  $\$d\$$ -dimensional model by replacing stochastic gradient directions with stochastic differences of  $\$K+1\$$  perturbed loss evaluations per dataset (example) query. For both unbounded and bounded possibly nonconvex losses, we present the first generalization bounds for the ZoSS algorithm. These bounds coincide with those for SGD, and rather surprisingly are independent of  $\$d\$, \$K\$$  and the batch size  $\$m\$$ , under appropriate choices of a slightly decreased learning rate. For bounded nonconvex losses and a batch size  $\$m=1\$$ , we additionally show that both generalization error and learning rate are independent of  $\$d\$$  and  $\$K\$$ , and remain essentially the same as for the SGD, even for two function evaluations. Our results extensively extend and consistently recover established results for SGD in prior work, on both generalization bounds and corresponding learning rates. If additionally  $\$m=n\$$ , where  $\$n\$$  is the dataset size, we derive generalization guarantees for full-batch GD as well.

## [PDSketch: Integrated Domain Programming, Learning, and Planning](#)

- Jiayuan Mao · Tomás Lozano-Pérez · Josh Tenenbaum · Leslie Kaelbling
- abstract@[open-review](#): This paper studies a model learning and online planning approach towards building flexible and general robots. Specifically, we investigate how to exploit the locality and sparsity structures in the underlying environmental transition model to improve model generalization, data-efficiency, and runtime-efficiency. We present a new domain definition language, named PDSketch. It allows users to flexibly define high-level structures in the transition models, such as object and feature dependencies, in a way similar to how programmers use TensorFlow or PyTorch to specify kernel sizes and hidden dimensions of a convolutional neural network. The details of the transition model will be filled in by trainable neural networks. Based on the defined structures and learned parameters, PDSketch automatically generates domain-independent planning heuristics without additional training. The derived heuristics accelerate the performance-time planning for novel goals.

## [Contrastive Graph Structure Learning via Information Bottleneck for Recommendation](#)

- Chunyu Wei · Jian Liang · Di Liu · Fei Wang
- abstract@[open-review](#): Graph convolution networks (GCNs) for recommendations have emerged as an important research topic due to their ability to exploit higher-order neighbors. Despite their success, most of them suffer from the popularity bias brought by a small number of active users and popular items. Also, a real-world user-item bipartite graph contains many noisy interactions, which may hamper the sensitive GCNs. Graph contrastive learning show promising performance for solving the above challenges in recommender systems. Most existing works typically perform graph augmentation to create multiple views of the original graph by randomly dropping edges/nodes or relying on predefined rules, and these augmented views always serve as an auxiliary task by maximizing their correspondence. However, we argue that the graph structures generated from these vanilla approaches may be suboptimal, and maximizing their correspondence will force the representation to capture information irrelevant for the recommendation task. Here, we propose a Contrastive Graph Structure Learning via Information Bottleneck (CGI) for recommendation, which adaptively learns whether to drop an edge or node to obtain optimized graph structures in an end-to-end manner. Moreover, we innovatively introduce the Information Bottleneck into the contrastive learning process to avoid capturing irrelevant information among different views and help enrich the final representation for recommendation. Extensive experiments on public datasets are provided to show that our model significantly outperforms strong baselines.

## [Self-Supervised Fair Representation Learning without Demographics](#)

- Junyi Chai · Xiaoqian Wang
- abstract@[open-review](#): Fairness has become an important topic in machine learning. Generally, most literature on fairness assumes that the sensitive information, such as gender or race, is present in the training set, and uses this information to mitigate bias. However, due to practical concerns like privacy and regulation, applications of these methods are restricted. Also, although much of the literature studies supervised learning, in many real-world scenarios, we want to utilize the large unlabelled dataset to improve the model's accuracy. Can we improve fair classification without sensitive information and without labels? To tackle the problem, in this paper, we propose a novel reweighing-based contrastive learning method. The goal of our method is to learn a generally fair representation without observing sensitive attributes. Our method assigns weights to training samples per iteration based on their gradient directions relative to the validation samples such that the average top-k validation loss is minimized. Compared with past fairness methods without demographics, our method is built on fully unsupervised training data and requires only a small labelled validation set. We provide rigorous theoretical proof of the convergence of our model. Experimental results show that our proposed method achieves better or comparable performance than state-of-the-art methods on three datasets in terms of accuracy and several fairness metrics.

## [Graph Self-supervised Learning with Accurate Discrepancy Learning](#)

- Dongki Kim · Jinheon Baek · Sung Ju Hwang
- abstract@[open-review](#): Self-supervised learning of graph neural networks (GNNs) aims to learn an accurate representation of the graphs in an unsupervised manner, to obtain transferable representations of them for diverse downstream tasks. Predictive learning and contrastive learning are the two most prevalent approaches for graph self-supervised learning. However, they have their own drawbacks. While the predictive learning methods can learn the contextual relationships between neighboring nodes and edges, they cannot learn global graph-level similarities. Contrastive learning, while it can learn global graph-level similarities, its objective to maximize the similarity between two differently perturbed graphs may result in representations that cannot discriminate two similar graphs with different properties. To tackle such limitations, we propose a framework that aims to learn the exact discrepancy between the original and the perturbed graphs, coined as Discrepancy-based Self-supervised LeArning (D-SLA). Specifically, we create multiple perturbations of the given graph with varying degrees of similarity, and train the model to predict whether each graph is the original graph or the perturbed one. Moreover, we further aim to accurately capture the amount of discrepancy for each perturbed graph using the graph edit distance. We validate our D-SLA on various graph-related downstream tasks, including molecular property prediction, protein function prediction, and link prediction tasks, on which ours largely outperforms relevant baselines.

## [Sampling from Log-Concave Distributions with Infinity-Distance Guarantees](#)

- Oren Mangoubi · Nisheeth Vishnoi

- abstract@[open-review](#): For a  $d$ -dimensional log-concave distribution  $\pi(\theta) \propto e^{-f(\theta)}$  constrained to a convex body  $K$ , the problem of outputting samples from a distribution  $\nu$  which is  $\epsilon$ -close in infinity-distance  $\sup_{\theta \in K} |\log \frac{\nu(\theta)}{\pi(\theta)}|$  to  $\pi$  arises in differentially private optimization. While sampling within total-variation distance  $\epsilon$  of  $\pi$  can be done by algorithms whose runtime depends polylogarithmically on  $\frac{1}{\epsilon}$ , prior algorithms for sampling in  $\epsilon$  infinity distance have runtime bounds that depend polynomially on  $\frac{1}{\epsilon}$ . We bridge this gap by presenting an algorithm that outputs a point  $\epsilon$ -close to  $\pi$  in infinity distance that requires at most  $\mathrm{poly}(\log \frac{1}{\epsilon}, d)$  calls to a membership oracle for  $K$  and evaluation oracle for  $f$ , when  $f$  is Lipschitz. Our approach departs from prior works that construct Markov chains on a  $\frac{1}{\epsilon^2}$ -discretization of  $K$  to achieve a sample with  $\epsilon$  infinity-distance error, and present a method to directly convert continuous samples from  $K$  with total-variation bounds to samples with infinity bounds. This approach also allows us to obtain an improvement on the dimension  $d$  in the running time for the problem of sampling from a log-concave distribution on polytopes  $K$  with infinity distance  $\epsilon$ , by plugging in TV-distance running time bounds for the Dikin Walk Markov chain.

## [\(Nearly\) All Cardinality Estimators Are Differentially Private](#)

- Charlie Dickens · Justin Thaler · Daniel Ting
- abstract@[open-review](#): We consider privacy in the context of streaming algorithms for cardinality estimation. We show that a large class of algorithms all satisfy  $\epsilon$ -differential privacy, so long as (a) the algorithm is combined with a simple down-sampling procedure, and (b) the input stream cardinality is  $\Omega(k/\epsilon)$ . Here,  $k$  is a certain parameter of the sketch that is always at most the sketch size in bits, but is typically much smaller. We also show that, even with no modification, algorithms in our class satisfy  $(\epsilon, \delta)$ -differential privacy, where  $\delta$  falls exponentially with the stream cardinality. Our analysis applies to essentially all popular cardinality estimation algorithms, and substantially generalizes and tightens privacy bounds from earlier works. Our approach is faster and exhibits a better utility-space tradeoff than prior art.

## [Torsional Diffusion for Molecular Conformer Generation](#)

- Bowen Jing · Gabriele Corso · Jeffrey Chang · Regina Barzilay · Tommi Jaakkola
- abstract@[open-review](#): Molecular conformer generation is a fundamental task in computational chemistry. Several machine learning approaches have been developed, but none have outperformed state-of-the-art cheminformatics methods. We propose torsional diffusion, a novel diffusion framework that operates on the space of torsion angles via a diffusion process on the hypertorus and an extrinsic-to-intrinsic score model. On a standard benchmark of drug-like molecules, torsional diffusion generates superior conformer ensembles compared to machine learning and cheminformatics methods in terms of both RMSD and chemical properties, and is orders of magnitude faster than competing diffusion-based models. Moreover, our model provides exact likelihoods, which we employ to build the first generalizable Boltzmann generator.

## [Provably expressive temporal graph networks](#)

- Amauri Souza · Diego Mesquita · Samuel Kaski · Vikas Garg
- abstract@[open-review](#): Temporal graph networks (TGNs) have gained prominence as models for embedding dynamic interactions, but little is known about their theoretical underpinnings. We establish fundamental results about the representational power and limits of the two main categories of TGNs: WA-TGNs that aggregate temporal walks, and MP-TGNs that augment local message passing with (recurrent) memory modules. Specifically, novel constructions reveal the inadequacy of MP-TGNs and WA-TGNs, proving that neither category subsumes the other. We extend the 1-WL (Weisfeiler-Leman) test to temporal graphs, and show that the most powerful MP-TGNs should use injective updates, as in this case they become as expressive as the temporal WL. Moreover, we elucidate that sufficiently deep MP-TGNs cannot benefit from memory, and MP-TGNs fail to compute graph properties such as girth. These theoretical insights lead us to introduce PINT --- a method provably more expressive than MP-TGN, WA-TGN, and temporal WL. Our experiments demonstrate that PINT outperforms existing TGNs on several real-world benchmarks.

## [Matrix Multiplicative Weights Updates in Quantum Zero-Sum Games: Conservation Laws & Recurrence](#)

- Rahul Jain · Georgios Piliouras · Ryann Sim
- abstract@[open-review](#): Recent advances in quantum computing and in particular, the introduction of quantum GANs, have led to increased interest in quantum zero-sum game theory, extending the scope of learning algorithms for classical games into the quantum realm. In this paper, we focus on learning in quantum zero-sum games under Matrix Multiplicative Weights Update (a generalization of the multiplicative weights update method) and its continuous analogue, Quantum Replicator Dynamics. When each player selects their state according to quantum replicator dynamics, we show that the system exhibits conservation laws in a quantum-information theoretic sense. Moreover, we show that the system exhibits Poincare recurrence, meaning that almost all orbits return arbitrarily close to their initial conditions infinitely often. Our analysis generalizes previous results in the case of classical games.

## [Empirical Phase Diagram for Three-layer Neural Networks with Infinite Width](#)

- Hanxu Zhou · Zhou Qixuan · Zhenyuan Jin · Tao Luo · Yaoyu Zhang · Zhi-Qin Xu
- abstract@[open-review](#): Substantial work indicates that the dynamics of neural networks (NNs) is closely related to their initialization of parameters. Inspired by the phase diagram for two-layer ReLU NNs with infinite width (Luo et al., 2021), we make a step towards drawing a phase diagram for three-layer ReLU NNs with infinite width. First, we derive a normalized gradient flow for three-layer ReLU NNs and obtain two key independent quantities to distinguish different dynamical regimes for common initialization methods. With carefully designed experiments and a large computation cost, for both synthetic datasets and real datasets, we find that the dynamics of each layer also could be divided into a linear regime and a condensed regime, separated by a critical regime. The criteria is the relative change of input weights (the input weight of a hidden neuron consists of the weight from its input layer to the hidden neuron and its bias term) as the width approaches infinity during the training, which tends to  $0$ ,  $+\infty$  and  $O(1)$ , respectively. In addition, we also demonstrate that different layers can lie in different dynamical regimes in a training process within a deep NN. In the condensed regime, we also observe the condensation of weights in isolated orientations with low complexity. Through experiments under three-layer condition, our phase diagram suggests a complicated dynamical regimes consisting of three possible regimes, together with their mixture, for deep NNs and provides a guidance for studying deep NNs in different initialization regimes, which reveals the possibility of completely different dynamics emerging within a deep NN for its different layers.

## [Adaptive Multi-stage Density Ratio Estimation for Learning Latent Space Energy-based Model](#)

- Zhisheng Xiao · Tian Han
- abstract@[open-review](#): This paper studies the fundamental problem of learning energy-based model (EBM) in the latent space of the generator model. Learning such prior model typically requires running costly Markov Chain Monte Carlo (MCMC). Instead, we propose to use noise contrastive estimation (NCE) to discriminatively learn the EBM through density ratio estimation between the latent prior density and latent posterior density. However, the NCE typically fails to accurately estimate such density ratio given large gap between two densities. To effectively tackle this issue and further learn more expressive prior model, we develop the adaptive multi-stage density ratio estimation which breaks the estimation into multiple stages and learn different stages of density ratio sequentially and adaptively. The latent prior model can be gradually learned using ratio estimated in previous stage so that the final latent space EBM prior can be naturally formed by product of ratios in different stages. The proposed method enables informative and much sharper prior

than existing baselines, and can be trained efficiently. Our experiments demonstrate strong performances in terms of image generation and reconstruction as well as anomaly detection.

## [On the Symmetries of Deep Learning Models and their Internal Representations](#)

- Charles Godfrey · Davis Brown · Tegan Emerson · Henry Kvinge
- abstract@[open-review](#): Symmetry has been a fundamental tool in the exploration of a broad range of complex systems. In machine learning, symmetry has been explored in both models and data. In this paper we seek to connect the symmetries arising from the architecture of a family of models with the symmetries of that family's internal representation of data. We do this by calculating a set of fundamental symmetry groups, which we call the intertwiner groups of the model. Each of these arises from a particular nonlinear layer of the model and different nonlinearities result in different symmetry groups. These groups change the weights of a model in such a way that the underlying function that the model represents remains constant but the internal representations of data inside the model may change. We connect intertwiner groups to a model's internal representations of data through a range of experiments that probe similarities between hidden states across models with the same architecture. Our work suggests that the symmetries of a network are propagated into the symmetries in that network's representation of data, providing us with a better understanding of how architecture affects the learning and prediction process. Finally, we speculate that for ReLU networks, the intertwiner groups may provide a justification for the common practice of concentrating model interpretability exploration on the activation basis in hidden layers rather than arbitrary linear combinations thereof.

## [Brownian Noise Reduction: Maximizing Privacy Subject to Accuracy Constraints](#)

- Justin Whitehouse · Aaditya Ramdas · Steven Wu · Ryan Rogers
- abstract@[open-review](#): There is a disconnect between how researchers and practitioners handle privacy-utility tradeoffs. Researchers primarily operate from a privacy first perspective, setting strict privacy requirements and minimizing risk subject to these constraints. Practitioners often desire an accuracy first perspective, possibly satisfied with the greatest privacy they can get subject to obtaining sufficiently small error. Ligett et al. have introduced a "noise reduction" algorithm to address the latter perspective. The authors show that by adding correlated Laplace noise and progressively reducing it on demand, it is possible to produce a sequence of increasingly accurate estimates of a private parameter and only pay a privacy cost for the least noisy iterate released. In this work, we generalize noise reduction to the setting of Gaussian noise, introducing the Brownian mechanism. The Brownian mechanism works by first adding Gaussian noise of high variance corresponding to the final point of a simulated Brownian motion. Then, at the practitioner's discretion, noise is gradually decreased by tracing back along the Brownian path to an earlier time. Our mechanism is more naturally applicable to the common setting of bounded  $\ell_2$ -sensitivity, empirically outperforms existing work on common statistical tasks, and provides customizable control of privacy loss over the entire interaction with the practitioner. We complement our Brownian mechanism with ReducedAboveThreshold, a generalization of the classical AboveThreshold algorithm that provides adaptive privacy guarantees. Overall, our results demonstrate that one can meet utility constraints while still maintaining strong levels of privacy.

## [Is Integer Arithmetic Enough for Deep Learning Training?](#)

- Alireza Ghaffari · Marzieh S. Tahaei · Mohammadreza Tayaranian · Masoud Asgharian · Vahid Partovi Nia
- abstract@[open-review](#): The ever-increasing computational complexity of deep learning models makes their training and deployment difficult on various cloud and edge platforms. Replacing floating-point arithmetic with low-bit integer arithmetic is a promising approach to save energy, memory footprint, and latency of deep learning models. As such, quantization has attracted the attention of researchers in recent years. However, using integer numbers to form a fully functional integer training pipeline including forward pass, back-propagation, and stochastic gradient descent is not studied in detail. Our empirical and mathematical results reveal that integer arithmetic seems to be enough to train deep learning models. Unlike recent proposals, instead of quantization, we directly switch the number representation of computations. Our novel training method forms a fully integer training pipeline that does not change the trajectory of the loss and accuracy compared to floating-point, nor does it need any special hyper-parameter tuning, distribution adjustment, or gradient clipping. Our experimental results show that our proposed method is effective in a wide variety of tasks such as classification (including vision transformers), object detection, and semantic segmentation.

## [Increasing the Scope as You Learn: Adaptive Bayesian Optimization in Nested Subspaces](#)

- Leonard Papenmeier · Matthias Poloczek · Luigi Nardi
- abstract@[open-review](#): Recent advances have extended the scope of Bayesian optimization (BO) to expensive-to-evaluate black-box functions with dozens of dimensions, aspiring to unlock impactful applications, for example, in the life sciences, neural architecture search, and robotics. However, a closer examination reveals that the state-of-the-art methods for high-dimensional Bayesian optimization (HDBO) suffer from degrading performance as the number of dimensions increases, or even risk failure if certain unverifiable assumptions are not met. This paper proposes BAxUS that leverages a novel family of nested random subspaces to adapt the space it optimizes over to the problem. This ensures high performance while removing the risk of failure, which we assert via theoretical guarantees. A comprehensive evaluation demonstrates that BAxUS achieves better results than the state-of-the-art methods for a broad set of applications.

## [Asymptotic Properties for Bayesian Neural Network in Besov Space](#)

- Kyeongwon Lee · Jaeyong Lee
- abstract@[open-review](#): Neural networks have shown great predictive power when dealing with various unstructured data such as images and natural languages. The Bayesian neural network captures the uncertainty of prediction by putting a prior distribution for the parameter of the model and computing the posterior distribution. In this paper, we show that the Bayesian neural network using spike-and-slab prior has consistency with nearly minimax convergence rate when the true regression function is in the Besov space. Even when the smoothness of the regression function is unknown the same posterior convergence rate holds and thus the spike and slab prior is adaptive to the smoothness of the regression function. We also consider the shrinkage prior, which is more feasible than other priors, and show that it has the same convergence rate. In other words, we propose a practical Bayesian neural network with guaranteed asymptotic properties.

## [Class-Dependent Label-Noise Learning with Cycle-Consistency Regularization](#)

- De Cheng · Yixiong Ning · Nannan Wang · Xinbo Gao · Heng Yang · Yuxuan Du · Bo Han · Tongliang Liu
- abstract@[open-review](#): In label-noise learning, estimating the transition matrix plays an important role in building statistically consistent classifier. Current state-of-the-art consistent estimator for the transition matrix has been developed under the newly proposed sufficiently scattered assumption, through incorporating the minimum volume constraint of the transition matrix  $T$  into label-noise learning. To compute the volume of  $T$ , it heavily relies on the estimated noisy class posterior. However, the estimation error of the noisy class posterior could usually be large as deep learning methods tend to easily overfit the noisy labels. Then, directly minimizing the volume of such obtained  $T$  could lead the transition matrix to be poorly estimated. Therefore, how to reduce the side-effects of the inaccurate noisy class posterior has become the bottleneck of such method. In this paper, we creatively propose to estimate the transition matrix under the forward-backward cycle-consistency regularization, of which we have greatly reduced the dependency of estimating the transition matrix  $T$  on the noisy class posterior. We show that the cycle-consistency regularization helps to minimize the volume of the transition matrix  $T$  indirectly without exploiting the estimated noisy class posterior, which could further encourage the estimated transition matrix  $T$  to

converge to its optimal solution. Extensive experimental results consistently justify the effectiveness of the proposed method, on reducing the estimation error of the transition matrix and greatly boosting the classification performance.

## [MOVE: Unsupervised Movable Object Segmentation and Detection](#)

- Adam Bielski · Paolo Favaro
- abstract@[open-review](#): We introduce MOVE, a novel method to segment objects without any form of supervision. MOVE exploits the fact that foreground objects can be shifted locally relative to their initial position and result in realistic (undistorted) new images. This property allows us to train a segmentation model that achieves state of the art (SotA) performance on several evaluation datasets for unsupervised salient object detection and segmentation. In unsupervised single object discovery, MOVE gives an average CorLoc improvement of 4.5% over the SotA, and in unsupervised class-agnostic object detection it gives a relative AP improvement of 30% on average. Our approach is built on top of self-supervised features (from DINO), an inpainting network (based on the Masked AutoEncoder) and adversarial training with a projected discriminator.

## [Offline Multi-Agent Reinforcement Learning with Knowledge Distillation](#)

- Wei-Cheng Tseng · Tsun-Hsuan Wang · Yen-Chen Lin · Phillip Isola
- abstract@[open-review](#): We introduce an offline multi-agent reinforcement learning (offline MARL) framework that utilizes previously collected data without additional online data collection. Our method reformulates offline MARL as a sequence modeling problem and thus builds on top of the simplicity and scalability of the Transformer architecture. In the fashion of centralized training and decentralized execution, we propose to first train a teacher policy as if the MARL dataset is generated by a single agent. After the teacher policy has identified and recombined the "good" behavior in the dataset, we create separate student policies and distill not only the teacher policy's features but also its structural relations among different agents' features to student policies. Despite its simplicity, the proposed method outperforms state-of-the-art model-free offline MARL baselines while being more robust to demonstration's quality on several environments.

## [Finite Sample Analysis Of Dynamic Regression Parameter Learning](#)

- Mark Kozdoba · Edward Moroshko · Shie Mannor · Yacov Crammer
- abstract@[open-review](#): We consider the dynamic linear regression problem, where the predictor vector may vary with time. This problem can be modeled as a linear dynamical system, with non-constant observation operator, where the parameters that need to be learned are the variance of both the process noise and the observation noise. While variance estimation for dynamic regression is a natural problem, with a variety of applications, existing approaches to this problem either lack guarantees altogether, or only have asymptotic guarantees without explicit rates. In particular, existing literature does not provide any clues to the following fundamental question: In terms of data characteristics, what does the convergence rate depend on? In this paper we study the global system operator -- the operator that maps the noise vectors to the output. We obtain estimates on its spectrum, and as a result derive the first known variance estimators with finite sample complexity guarantees. The proposed bounds depend on the shape of a certain spectrum related to the system operator, and thus provide the first known explicit geometric parameter of the data that can be used to bound estimation errors. In addition, the results hold for arbitrary sub Gaussian distributions of noise terms. We evaluate the approach on synthetic and real-world benchmarks.

## [Variable-rate hierarchical CPC leads to acoustic unit discovery in speech](#)

- Santiago Cuervo · Adrian Lancucki · Ricard Marxer · Paweł Rychlikowski · Jan Chorowski
- abstract@[open-review](#): The success of deep learning comes from its ability to capture the hierarchical structure of data by learning high-level representations defined in terms of low-level ones. In this paper we explore self-supervised learning of hierarchical representations of speech by applying multiple levels of Contrastive Predictive Coding (CPC). We observe that simply stacking two CPC models does not yield significant improvements over single-level architectures. Inspired by the fact that speech is often described as a sequence of discrete units unevenly distributed in time, we propose a model in which the output of a low-level CPC module is non-uniformly downsampled to directly minimize the loss of a high-level CPC module. The latter is designed to also enforce a prior of separability and discreteness in its representations by enforcing dissimilarity of successive high-level representations through focused negative sampling, and by quantization of the prediction targets. Accounting for the structure of the speech signal improves upon single-level CPC features and enhances the disentanglement of the learned representations, as measured by downstream speech recognition tasks, while resulting in a meaningful segmentation of the signal that closely resembles phone boundaries.

## [Heterogeneous Skill Learning for Multi-agent Tasks](#)

- Yuntao Liu · Yuan Li · Xinhai Xu · Yong Dou · Donghong Liu
- abstract@[open-review](#): Heterogeneous behaviours are widespread in many multi-agent tasks, which have not been paid much attention in the community of multi-agent reinforcement learning. It would be a key factor for improving the learning performance to efficiently characterize and automatically find heterogeneous behaviours. In this paper, we introduce the concept of the skill to explore the ability of heterogeneous behaviours. We propose a novel skill-based multi-agent reinforcement learning framework to enable agents to master diverse skills. Specifically, our framework consists of the skill representation mechanism, the skill selector and the skill-based policy learning mechanism. We design an auto-encoder model to generate the latent variable as the skill representation by incorporating the environment information, which ensures the distinguishable of agents for skill selection and the discriminability for the skill learning. With the representation, a skill selection mechanism is invented to realize the assignment from agents to skills. Meanwhile, diverse skill-based policies are generated through a novel skill-based policy learning method. To promote efficient skill discovery, a mutual information based intrinsic reward function is constructed. Empirical results show that our framework obtains the best performance on three challenging benchmarks, i.e., StarCraft II micromanagement tasks, Google Research Football and GoBigger, over state-of-the-art MARL methods.

## [FedAvg with Fine Tuning: Local Updates Lead to Representation Learning](#)

- Liam Collins · Hamed Hassani · Aryan Mokhtari · Sanjay Shakkottai
- abstract@[open-review](#): The Federated Averaging (FedAvg) algorithm, which consists of alternating between a few local stochastic gradient updates at client nodes, followed by a model averaging update at the server, is perhaps the most commonly used method in Federated Learning. Notwithstanding its simplicity, several empirical studies have illustrated that the output model of FedAvg, after a few fine-tuning steps, leads to a model that generalizes well to new unseen tasks. This surprising performance of such a simple method, however, is not fully understood from a theoretical point of view. In this paper, we formally investigate this phenomenon in the multi-task linear representation setting. We show that the reason behind generalizability of the FedAvg's output is its power in learning the common data representation among the clients' tasks, by leveraging the diversity among client data distributions via local updates. We formally establish the iteration complexity required by the clients for proving such result in the setting where the underlying shared representation is a linear map. To the best of our knowledge, this is the first such result for any setting. We also provide empirical evidence demonstrating FedAvg's representation learning ability in federated image classification with heterogeneous data.

## [NeuroSchedule: A Novel Effective GNN-based Scheduling Method for High-level Synthesis](#)

- Jun Zeng · Mingyang Kou · Hailong Yao

- abstract@[open-review](#): High-level synthesis (HLS) is widely used for transferring behavior-level specifications into circuit-level implementations. As a critical step in HLS, scheduling arranges the execution order of operations for enhanced performance. However, existing scheduling methods suffer from either exponential runtime or poor quality of solutions. This paper proposes NeuroSchedule, an efficient and effective GNN-based scheduling method called NeuroSchedule, with both fast runtime and enhanced solution quality. Major features are as follows: (1) The learning problem for HLS scheduling is formulated for the first time, and a new machine learning framework is proposed. (2) Pre-training models are adopted to further enhance the scalability for various scheduling problems with different settings. Experimental results show that NeuroSchedule obtains near-optimal solutions while achieving more than 50,000x improvement in runtime compared with the ILP-based scheduling method. At the same time, NeuroSchedule improves the scheduling results by 6.10% on average compared with state-of-the-art entropy-directed method. To the best of our knowledge, this is the first GNN-based scheduling method for HLS.

## [FourierNets enable the design of highly non-local optical encoders for computational imaging](#)

- Diptodip Deb · Zhenfei Jiao · Ruth Sims · Alex Chen · Michael Broxton · Misha B Ahrens · Kaspar Podgorski · Srinivas C Turaga
- abstract@[open-review](#): Differentiable simulations of optical systems can be combined with deep learning-based reconstruction networks to enable high performance computational imaging via end-to-end (E2E) optimization of both the optical encoder and the deep learning decoder. This has enabled imaging applications such as 3D localization microscopy, depth estimation, and lensless photography via the optimization of local optical encoders. More challenging computational imaging applications, such as 3D snapshot microscopy which compresses 3D volumes into single 2D images, require a highly non-local optical encoder. We show that existing deep network decoders have a locality bias which prevents the optimization of such highly non-local optical encoders. We address this with a decoder based on a shallow neural network architecture using global kernel Fourier convolutional neural networks (FourierNets). We show that FourierNets surpass existing deep network based decoders at reconstructing photographs captured by the highly non-local DiffuserCam optical encoder. Further, we show that FourierNets enable E2E optimization of highly non-local optical encoders for 3D snapshot microscopy. By combining FourierNets with a large-scale multi-GPU differentiable optical simulation, we are able to optimize non-local optical encoders 170\$times\$ to 7372\$times\$ larger than prior state of the art, and demonstrate the potential for ROI-type specific optical encoding with a programmable microscope.

## [Debiasing Graph Neural Networks via Learning Disentangled Causal Substructure](#)

- Shaohua Fan · Xiao Wang · Yanhu Mo · Chuan Shi · Jian Tang
- abstract@[open-review](#): Most Graph Neural Networks (GNNs) predict the labels of unseen graphs by learning the correlation between the input graphs and labels. However, by presenting a graph classification investigation on the training graphs with severe bias, surprisingly, we discover that GNNs always tend to explore the spurious correlations to make decision, even if the causal correlation always exists. This implies that existing GNNs trained on such biased datasets will suffer from poor generalization capability. By analyzing this problem in a causal view, we find that disentangling and decorrelating the causal and bias latent variables from the biased graphs are both crucial for debiasing. Inspiring by this, we propose a general disentangled GNN framework to learn the causal substructure and bias substructure, respectively. Particularly, we design a parameterized edge mask generator to explicitly split the input graph into causal and bias subgraphs. Then two GNN modules supervised by causal/bias-aware loss functions respectively are trained to encode causal and bias subgraphs into their corresponding representations. With the disentangled representations, we synthesize the counterfactual unbiased training samples to further decorrelate causal and bias variables. Moreover, to better benchmark the severe bias problem, we construct three new graph datasets, which have controllable bias degrees and are easier to visualize and explain. Experimental results well demonstrate that our approach achieves superior generalization performance over existing baselines. Furthermore, owing to the learned edge mask, the proposed model has appealing interpretability and transferability.

## [Learning to Constrain Policy Optimization with Virtual Trust Region](#)

- Thai Hung Le · Thommen Karimpanal George · Majid Abdolshah · Dung Nguyen · Kien Do · Sunil Gupta · Svetha Venkatesh
- abstract@[open-review](#): We introduce a constrained optimization method for policy gradient reinforcement learning, which uses two trust regions to regulate each policy update. In addition to using the proximity of one single old policy as the first trust region as done by prior works, we propose forming a second trust region by constructing another virtual policy that represents a wide range of past policies. We then enforce the new policy to stay closer to the virtual policy, which is beneficial if the old policy performs poorly. We propose a mechanism to automatically build the virtual policy from a memory buffer of past policies, providing a new capability for dynamically selecting appropriate trust regions during the optimization process. Our proposed method, dubbed Memory-Constrained Policy Optimization (MCPO), is examined in diverse environments, including robotic locomotion control, navigation with sparse rewards and Atari games, consistently demonstrating competitive performance against recent on-policy constrained policy gradient methods.

## [Quantifying Statistical Significance of Neural Network-based Image Segmentation by Selective Inference](#)

- Vo Nguyen Le Duy · Shogo Iwazaki · Ichiro Takeuchi
- abstract@[open-review](#): Although a vast body of literature relates to image segmentation methods that use deep neural networks (DNNs), less attention has been paid to assessing the statistical reliability of segmentation results. In this study, we interpret the segmentation results as hypotheses driven by DNN (called DNN-driven hypotheses) and propose a method to quantify the reliability of these hypotheses within a statistical hypothesis testing framework. To this end, we introduce a conditional selective inference (SI) framework---a new statistical inference framework for data-driven hypotheses that has recently received considerable attention---to compute exact (non-asymptotic) valid p-values for the segmentation results. To use the conditional SI framework for DNN-based segmentation, we develop a new SI algorithm based on the homotopy method, which enables us to derive the exact (non-asymptotic) sampling distribution of DNN-driven hypothesis. We conduct several experiments to demonstrate the performance of the proposed method.

## [Beyond L1: Faster and Better Sparse Models with skglm](#)

- Quentin Bertrand · Quentin Klopfenstein · Pierre-Antoine Bannier · Gauthier Gidel · Mathurin Massias
- abstract@[open-review](#): We propose a new fast algorithm to estimate any sparse generalized linear model with convex or non-convex separable penalties. Our algorithm is able to solve problems with millions of samples and features in seconds, by relying on coordinate descent, working sets and Anderson acceleration. It handles previously unaddressed models, and is extensively shown to improve state-of-art algorithms. We provide a flexible, scikit-learn compatible package, which easily handles customized datafits and penalties.

## [Brain Network Transformer](#)

- Xuan Kan · Wei Dai · Hejie Cui · Zilong Zhang · Ying Guo · Carl Yang
- abstract@[open-review](#): Human brains are commonly modeled as networks of Regions of Interest (ROIs) and their connections for the understanding of brain functions and mental disorders. Recently, Transformer-based models have been studied over different types of data, including graphs, shown to bring performance gains widely. In this work, we study Transformer-based models for brain network analysis. Driven by the unique properties of data, we model brain networks as graphs with nodes of fixed size and order, which allows us to (1) use connection profiles as node features to provide natural and low-cost positional information and (2) learn pair-wise connection strengths among ROIs with efficient attention weights across individuals that are predictive towards downstream analysis tasks. Moreover, we propose an Orthonormal Clustering Readout operation based on self-supervised soft clustering and orthonormal projection. This design accounts for the underlying functional modules that determine similar behaviors among groups of

ROIs, leading to distinguishable cluster-aware node embeddings and informative graph embeddings. Finally, we re-standardize the evaluation pipeline on the only one publicly available large-scale brain network dataset of ABIDE, to enable meaningful comparison of different models. Experiment results show clear improvements of our proposed Brain Network Transformer on both the public ABIDE and our restricted ABCD datasets. The implementation is available at <https://anonymous.4open.science/r/BrainTransformer>.

## [Towards Practical Computation of Singular Values of Convolutional Layers](#)

- Alexandra Senderovich · Ekaterina Bulatova · Anton Obukhov · Maxim Rakuba
- abstract@[open-review](#): In general, convolutional neural networks (CNNs) are easy to train, but their essential properties, such as generalization error and adversarial robustness, are hard to control. Recent research demonstrated that singular values of convolutional layers significantly affect such elusive properties and offered several methods for controlling them. Nevertheless, these methods present a significant computational challenge or resort to coarse approximations. In this paper, we offer a principled approach to alleviating constraints of the prior art at the expense of an insignificant reduction in layer expressivity. Our method is based on the tensor train decomposition; it retains control over the actual singular values of convolutional mappings while providing structurally sparse and hardware-friendly representation. We demonstrate the improved properties of modern CNNs with our method and analyze its impact on the model performance, calibration, and adversarial robustness.

## [EvenNet: Ignoring Odd-Hop Neighbors Improves Robustness of Graph Neural Networks](#)

- Runlin Lei · Zhen Wang · Yaliang Li · Bolin Ding · Zhewei Wei
- abstract@[open-review](#): Graph Neural Networks (GNNs) have received extensive research attention for their promising performance in graph machine learning. Despite their extraordinary predictive accuracy, existing approaches, such as GCN and GPRGNN, are not robust in the face of homophily changes on test graphs, rendering these models vulnerable to graph structural attacks and with limited capacity in generalizing to graphs of varied homophily levels. Although many methods have been proposed to improve the robustness of GNN models, most of these techniques are restricted to the spatial domain and employ complicated defense mechanisms, such as learning new graph structures or calculating edge attentions. In this paper, we study the problem of designing simple and robust GNN models in the spectral domain. We propose EvenNet, a spectral GNN corresponding to an even-polynomial graph filter. Based on our theoretical analysis in both spatial and spectral domains, we demonstrate that EvenNet outperforms full-order models in generalizing across homophilic and heterophilic graphs, implying that ignoring odd-hop neighbors improves the robustness of GNNs. We conduct experiments on both synthetic and real-world datasets to demonstrate the effectiveness of EvenNet. Notably, EvenNet outperforms existing defense models against structural attacks without introducing additional computational costs and maintains competitiveness in traditional node classification tasks on homophilic and heterophilic graphs.

## [Learning White Noises in Neural Stochastic Differential Equations](#)

- Anh Tong Hoang · Thanh Nguyen-Tang · Toan Tran · Jaesik Choi
- abstract@[open-review](#): Differential equations play important roles in modeling complex physical systems. Recent advances present interesting research directions by combining differential equations with neural networks. By including noise, stochastic differential equations (SDEs) allows us to model data with uncertainty and measure imprecision. There are many variants of noises known to exist in many real-world data. For example, previously white noises are idealized and induced by Brownian motions. Nevertheless, there is a lack of machine learning models that can handle such noises. In this paper, we introduce a generalized white noise to existing models and propose an efficient approximation of noise sample paths based on classical integration methods and sparse Gaussian process. Our experimental results demonstrate that the proposed model can capture noise characteristics such as continuity from various time series data, therefore improving model fittings over existing models. We examine how we can apply our approach to score-based generative models, showing that there exists a case of our generalized noise resulting in a better image generation measure.

## [Enhancing and Scaling Cross-Modality Alignment for Contrastive Multimodal Pre-Training via Gradient Harmonization](#)

- Junru Wu · Yi Liang · feng han · Hassan Akbari · Zhangyang Wang · Cong Yu
- abstract@[open-review](#): Self-supervised pre-training recently demonstrates success on large-scale multimodal data, and state-of-the-art contrastive methods often enforce the feature consistency from cross-modality inputs, such as video/audio or video/text pairs. Despite its convenience to formulate and leverage in practice, such cross-modality alignment (CMA) is only a weak and noisy supervision, since two modalities can be semantically misaligned even if they are temporally aligned. For example, even in the (often adopted) instructional videos, a speaker can sometimes refer to something that is not visually present in the current frame; and the semantic misalignment would only be more unpredictable for the raw videos collected from unconstrained internet sources. We conjecture that might cause conflicts and biases among modalities, and may hence prohibit CMA from scaling up to training with larger and more heterogeneous data. This paper first verifies our conjecture by observing that, even in the latest VATT pre-training using only narrated videos, there exist strong gradient conflicts between different CMA losses within the same sample triplet (video, audio, text), indicating them as the noisy source of supervision. We then propose to harmonize such gradients during pre-training, via two techniques: (i) cross-modality gradient realignment: modifying different CMA loss gradients for one sample triplet, so that their gradient directions are in more agreement; and (ii) gradient-based curriculum learning: leveraging the gradient conflict information on an indicator of sample noisiness, to develop a curriculum learning strategy to prioritize training with less noisy sample triplets. Applying those gradient harmonization techniques to pre-training VATT on the HowTo100M dataset, we consistently improve its performance on different downstream tasks. Moreover, we are able to scale VATT pre-training to more complicated non-narrative YouTube8M dataset to further improve the state-of-the-arts.

## [VAEL: Bridging Variational Autoencoders and Probabilistic Logic Programming](#)

- Eleonora Misino · Giuseppe Marra · Emanuele Sansone
- abstract@[open-review](#): We present VAEL, a neuro-symbolic generative model integrating variational autoencoders (VAE) with the reasoning capabilities of probabilistic logic (L) programming. Besides standard latent subsymbolic variables, our model exploits a probabilistic logic program to define a further structured representation, which is used for logical reasoning. The entire process is end-to-end differentiable. Once trained, VAEL can solve new unseen generation tasks by (i) leveraging the previously acquired knowledge encoded in the neural component and (ii) exploiting new logical programs on the structured latent space. Our experiments provide support on the benefits of this neuro-symbolic integration both in terms of task generalization and data efficiency. To the best of our knowledge, this work is the first to propose a general-purpose end-to-end framework integrating probabilistic logic programming into a deep generative model.

## [Generalization Bounds with Minimal Dependency on Hypothesis Class via Distributionally Robust Optimization](#)

- Yibo Zeng · Henry Lam
- abstract@[open-review](#): Established approaches to obtain generalization bounds in data-driven optimization and machine learning mostly build on solutions from empirical risk minimization (ERM), which depend crucially on the functional complexity of the hypothesis class. In this paper, we present an alternate route to obtain these bounds on the solution from distributionally robust optimization (DRO), a recent data-driven optimization framework based on worst-case analysis and the notion of ambiguity set to capture statistical uncertainty. In contrast to the hypothesis class complexity in ERM, our DRO bounds depend on the ambiguity set geometry and its compatibility with the true loss function. Notably, when using maximum mean discrepancy as a DRO distance metric, our analysis implies generalization bounds whose dependence on the hypothesis class appears the minimal possible: The bound depends solely on the true loss function, independent of any other candidates in the hypothesis class. To our best knowledge, it is the first generalization

bound of this type in the literature, and we hope our findings can open the door for a better understanding of DRO, especially its benefits on loss minimization and other machine learning applications.

## [Revisiting Injective Attacks on Recommender Systems](#)

- Haoyang LI · Shimin DI · Lei Chen
- abstract@[open-review](#): Recent studies have demonstrated that recommender systems (RecSys) are vulnerable to injective attacks. Given a limited fake user budget, attackers can inject fake users with carefully designed behaviors into the open platforms, making RecSys recommend a target item to more real users for profits. In this paper, we first revisit existing attackers and reveal that they suffer from the difficulty-agnostic and diversity-deficit issues. Existing attackers concentrate their efforts on difficult users who have low tendencies toward the target item, thus reducing their effectiveness. Moreover, they are incapable of affecting the target RecSys to recommend the target item to real users in a diverse manner, because their generated fake user behaviors are dominated by large communities. To alleviate these two issues, we propose a difficulty and diversity aware attacker, namely DADA. We design the difficulty-aware and diversity-aware objectives to enable easy users from various communities to contribute more weights when optimizing attackers. By incorporating these two objectives, the proposed attacker DADA can concentrate on easy users while also affecting a broader range of real users simultaneously, thereby boosting the effectiveness. Extensive experiments on three real-world datasets demonstrate the effectiveness of our proposed attacker.

## [Modeling the Machine Learning Multiverse](#)

- Samuel J. Bell · Onno Kampman · Jesse Dodge · Neil Lawrence
- abstract@[open-review](#): Amid mounting concern about the reliability and credibility of machine learning research, we present a principled framework for making robust and generalizable claims: the Multiverse Analysis. Our framework builds upon the Multiverse Analysis introduced in response to psychology's own reproducibility crisis. To efficiently explore high-dimensional and often continuous ML search spaces, we model the multiverse with a Gaussian Process surrogate and apply Bayesian experimental design. Our framework is designed to facilitate drawing robust scientific conclusions about model performance, and thus our approach focuses on exploration rather than conventional optimization. In the first of two case studies, we investigate disputed claims about the relative merit of adaptive optimizers. Second, we synthesize conflicting research on the effect of learning rate on the large batch training generalization gap. For the machine learning community, the Multiverse Analysis is a simple and effective technique for identifying robust claims, for increasing transparency, and a step toward improved reproducibility.

## [Uni-Perceiver-MoE: Learning Sparse Generalist Models with Conditional MoEs](#)

- Jinguo Zhu · Xizhou Zhu · Wenhui Wang · Xiaohua Wang · Hongsheng Li · Xiaogang Wang · Jifeng Dai
- abstract@[open-review](#): To build an artificial neural network model like a biological intelligence system, recent works have been unifying numerous tasks into a generalist model, which can process various tasks with shared parameters and do not have any task-specific modules. While generalist models achieve promising results on various benchmarks, they also have performance degradation on some tasks compared with task-specialized models. In this work, we find that interference among different tasks and modalities is the main factor to this phenomenon. To mitigate such interference, we introduce the Conditional Mixture of Experts (Conditional MoEs) to generalist models. Routing strategies under different levels of conditions are proposed to take both the training/inference cost and generalization ability into account. By incorporating the proposed Conditional MoEs, the recently proposed generalist model Uni-Perceiver can effectively mitigate the interference across tasks and modalities, and achieves new state-of-the-art results on a series of downstream tasks via prompt tuning on 1% of downstream data. Moreover, the introduction of Conditional MoEs still holds the generalization ability of generalist models to conduct zero-shot inference on new tasks, e.g., video-text retrieval and video caption. Code and pre-trained generalist models shall be released.

## [Variational Context Adjustment for Temporal Event Prediction under Distribution Shifts](#)

- Chenxiao Yang · Qitian Wu · Qingsong Wen · Zhiqiang Zhou · Liang Sun · Junchi Yan
- abstract@[open-review](#): The goal of event sequence modeling is to predict the next event based on a sequence of historical events, with applications to user behavior prediction, sequential recommendation and epidemic control. In practice, sequence learning models are trained with data collected at one time and need to handle new data in remote future, which requires models to handle temporal distribution shift from training to testing. In this paper, we first take a data-generating perspective to reveal a negative result that existing approaches with maximum likelihood estimation would fail for distribution shift due to the latent context confounder. Then we devise a new learning objective based on backdoor adjustment and further harness variational inference to make it tractable for sequence learning. On top of that, we propose a framework with hierarchical branching structures for learning context-specific representations. Comprehensive experiments on three sequence learning tasks demonstrate the effectiveness, applicability and scalability of our method with various off-the-shelf models as backbones.

## [How to talk to your model: Instructions, descriptions, and learning](#)

- Theodore Sumers · Robert Hawkins · Mark Ho · Tom Griffiths · Dylan Hadfield-Menell
- abstract@[open-review](#): From the earliest years of our lives, humans use language to express our beliefs and desires. Being able to talk to artificial agents about our preferences would thus fulfill a central goal of value alignment. Yet today, we lack computational models explaining such language use. To address this challenge, we formalize learning from language in a contextual bandit setting and ask how a human might communicate preferences over behaviors. We study two distinct types of language: instructions, which provide information about the desired policy, and descriptions, which provide information about the reward function. We show that the agent's degree of autonomy determines which form of language is optimal: instructions are better in low-autonomy settings, but descriptions are better when the agent will need to act independently. We then define a pragmatic listener agent that robustly infers the speaker's reward function by reasoning about how the speaker expresses themselves. We validate our models with a behavioral experiment, demonstrating that (1) our speaker model predicts human behavior, and (2) our pragmatic listener successfully recovers humans' reward functions. Finally, we show that this form of social learning can integrate with and reduce regret in traditional reinforcement learning. We hope these insights facilitate a shift from developing agents that obey language to agents that learn from it.

## [Geometric Distillation for Graph Networks](#)

- Chenxiao Yang · Qitian Wu · Junchi Yan
- abstract@[open-review](#): We study a new paradigm of knowledge transfer in the context of geometric deep learning, which aims at distilling knowledge from a teacher graph neural network (GNN) model trained on a large graph to a student GNN model operating on a smaller graph. To this end, we revisit the connection between thermodynamics and the behavior of GNN, based on which we propose Neural Heat Kernel (NHK) to encapsulate the geometric property of the underlying manifold concerning the architecture of GNN. A natural solution is derived by analysing and aligning NHKs on teacher and student models, dubbed as Geometric Knowledge Distillation. We develop non- and parametric instantiations and demonstrate their efficacy in various experimental settings for knowledge distillation regarding different types of privileged topological information and teacher-student schemes.

## [ComENet: Towards Complete and Efficient Message Passing for 3D Molecular Graphs](#)

- Limei Wang · Yi Liu · Yuchao Lin · Haoran Liu · Shuiwang Ji
- abstract@[open-review](#): Many real-world data can be modeled as 3D graphs, but learning representations that incorporates 3D information completely and efficiently is challenging. Existing methods either use partial 3D information, or suffer from excessive computational cost. To incorporate 3D information completely and efficiently, we propose a novel message passing scheme that operates within 1-hop neighborhood. Our method guarantees full completeness of 3D information on 3D graphs by achieving global and local completeness. Notably, we propose the important rotation angles to fulfill global completeness. Additionally, we show that our method is orders of magnitude faster than prior methods. We provide rigorous proof of completeness and analysis of time complexity for our methods. As molecules are in essence quantum systems, we build the \underline{com}plete and \underline{e}fficient graph neural network (ComENet) by combining quantum inspired basis functions and the proposed message passing scheme. Experimental results demonstrate the capability and efficiency of ComENet, especially on real-world datasets that are large in both numbers and sizes of graphs. Our code is publicly available as part of the DIG library (\url{https://github.com/divelab/DIG}).

## [Active Learning Through a Covering Lens](#)

- Ofer Yehuda · Avihu Dekel · Guy Hacohen · Daphna Weinshall
- abstract@[open-review](#): Deep active learning aims to reduce the annotation cost for deep neural networks, which are notoriously data-hungry. Until recently, deep active learning methods struggled in the low-budget regime, where only a small amount of samples are annotated. The situation has been alleviated by recent advances in self-supervised representation learning methods, which impart the geometry of the data representation with rich information about the points. Taking advantage of this progress, we study the problem of subset selection for annotation through a “covering” lens, proposing ProbCover -- a new active learning algorithm for the low budget regime, which seeks to maximize Probability Coverage. We describe a dual way to view our formulation, from which one can derive strategies suitable for the high budget regime of active learning, related to existing methods like Coreset. We conclude with extensive experiments, evaluating ProbCover in the low budget regime. We show that our principled active learning strategy improves the state-of-the-art in the low-budget regime in several image recognition benchmarks. This method is especially beneficial in semi-supervised settings, allowing state-of-the-art semi-supervised methods to achieve high accuracy with only a few labels.

## [Descent Steps of a Relation-Aware Energy Produce Heterogeneous Graph Neural Networks](#)

- Hongjoon Ahn · Yongyi Yang · Quan Gan · David P Wipf · Taesup Moon
- abstract@[open-review](#): Heterogeneous graph neural networks (GNNs) achieve strong performance on node classification tasks in a semi-supervised learning setting. However, as in the simpler homogeneous GNN case, message-passing-based heterogeneous GNNs may struggle to balance between resisting the oversmoothing occurring in deep models and capturing long-range dependencies in structured data. Moreover, the complexity of this trade-off is compounded in the heterogeneous graph case due to the disparate heterophily relationships between nodes of different types. To address these issues, we proposed a novel heterogeneous GNN architecture in which layers are derived from optimization steps that descend a novel relation-aware energy function. The corresponding minimizer is fully differentiable with respect to the energy function parameters, such that bilevel optimization can be applied to effectively learn a functional form whose minimum provides optimal node representations for subsequent classification tasks. In particular, this methodology allows us to model diverse heterophily relationships between different node types while avoiding oversmoothing effects. Experimental results on 8 heterogeneous graph benchmarks demonstrates that our proposed method can achieve competitive node classification accuracy.

## [Fair Rank Aggregation](#)

- Diptarka Chakraborty · Syamantak Das · Arindam Khan · Aditya Subramanian
- abstract@[open-review](#): Ranking algorithms find extensive usage in diverse areas such as web search, employment, college admission, voting, etc. The related rank aggregation problem deals with combining multiple rankings into a single aggregate ranking. However, algorithms for both these problems might be biased against some individuals or groups due to implicit prejudice or marginalization in the historical data. We study ranking and rank aggregation problems from a fairness or diversity perspective, where the candidates (to be ranked) may belong to different groups and each group should have a fair representation in the final ranking. We allow the designer to set the parameters that define fair representation. These parameters specify the allowed range for the number of candidates from a particular group in the top-\$k\$ positions of the ranking. We provide a linear time exact algorithm for finding the closest fair ranking for the Kendall-Tau metric under strong fairness, i.e., when the final ranking is fair for all reasonably large values of \$k\$. We also provide an exact algorithm for finding the closest fair ranking for the Ulam metric under strong fairness when there are only \$O(1)\$ number of groups. Our algorithms are simple and might be extendable to other relevant metrics. We also give a meta-algorithm for the general rank aggregation problem under the fairness framework. Surprisingly, this meta-algorithm works for any generalized mean objective (including center and median problems) and any fairness criteria. As a byproduct, we obtain 3-approximation algorithms for both center and median problems, under both Kendall Tau and Ulam metrics. Furthermore, using sophisticated techniques we obtain a \$(3-\epsilon)\$-approximation algorithm for the Ulam metric under strong fairness, for a constant \$\epsilon > 0\$.

## [One Positive Label is Sufficient: Single-Positive Multi-Label Learning with Label Enhancement](#)

- Ning Xu · Congyu Qiao · Jiaqi Lv · Xin Geng · Min-Ling Zhang
- abstract@[open-review](#): Multi-label learning (MLL) learns from the examples each associated with multiple labels simultaneously, where the high cost of annotating all relevant labels for each training example is challenging for real-world applications. To cope with the challenge, we investigate single-positive multi-label learning (SPMLL) where each example is annotated with only one relevant label and show that one can successfully learn a theoretically grounded multi-label classifier for the problem. In this paper, a novel SPMLL method named SMILE, i.e., Single-positive Multi-label learning with Label Enhancement, is proposed. Specifically, an unbiased risk estimator is derived, which could be guaranteed to approximately converge to the optimal risk minimizer of fully supervised learning and shows that one positive label of each instance is sufficient to train the predictive model. Then, the corresponding empirical risk estimator is established via recovering the latent soft label as a label enhancement process, where the posterior density of the latent soft labels is approximate to the variational Beta density parameterized by an inference model. Experiments on benchmark datasets validate the effectiveness of the proposed method.

## [Optimal Binary Classification Beyond Accuracy](#)

- Shashank Singh · Justin Khim
- abstract@[open-review](#): The vast majority of statistical theory on binary classification characterizes performance in terms of accuracy. However, accuracy is known in many cases to poorly reflect the practical consequences of classification error, most famously in imbalanced binary classification, where data are dominated by samples from one of two classes. The first part of this paper derives a novel generalization of the Bayes-optimal classifier from accuracy to any performance metric computed from the confusion matrix. Specifically, this result (a) demonstrates that stochastic classifiers sometimes outperform the best possible deterministic classifier and (b) removes an empirically unverifiable absolute continuity assumption that is poorly understood but pervades existing results. We then demonstrate how to use this generalized Bayes classifier to obtain regret bounds in terms of the error of estimating regression functions under uniform loss. Finally, we use these results to develop some of the first finite-sample statistical guarantees specific to imbalanced binary classification. Specifically, we demonstrate that optimal classification performance depends on properties of class imbalance, such as a novel notion called Uniform Class Imbalance, that have not previously been formalized. We further illustrate these contributions numerically in the case of \$k\$-nearest neighbor classification.

## [Consistent Sufficient Explanations and Minimal Local Rules for explaining the decision of any classifier or regressor](#)

- Salim I. Amoukou Â· Nicolas Brunel
- abstract@[open-review](#): To explain the decision of any regression and classification model, we extend the notion of probabilistic sufficient explanations (P-SE). For each instance, this approach selects the minimal subset of features that is sufficient to yield the same prediction with high probability, while removing other features. The crux of P-SE is to compute the conditional probability of maintaining the same prediction. Therefore, we introduce an accurate and fast estimator of this probability via random Forests for any data  $\{\boldsymbol{X}\}, Y$  and show its efficiency through a theoretical analysis of its consistency. As a consequence, we extend the P-SE to regression problems. In addition, we deal with non-binary features, without learning the distribution of  $\boldsymbol{X}$  nor having the model for making predictions. Finally, we introduce local rule-based explanations for regression/classification based on the P-SE and compare our approaches w.r.t other explainable AI methods. These methods are available as a Python Package.

## [PaCo: Parameter-Compositional Multi-task Reinforcement Learning](#)

- Lingfeng Sun Â· Haichao Zhang Â· Wei Xu Â· Masayoshi TOMIZUKA
- abstract@[open-review](#): The purpose of multi-task reinforcement learning (MTRL) is to train a single policy that can be applied to a set of different tasks. Sharing parameters allows us to take advantage of the similarities among tasks. However, the gaps between contents and difficulties of different tasks bring us challenges on both which tasks should share the parameters and what parameters should be shared, as well as the optimization challenges due to parameter sharing. In this work, we introduce a parameter-compositional approach (PaCo) as an attempt to address these challenges. In this framework, a policy subspace represented by a set of parameters is learned. Policies for all the single tasks lie in this subspace and can be composed by interpolating with the learned set. It allows not only flexible parameter sharing, but also a natural way to improve training. We demonstrate the state-of-the-art performance on Meta-World benchmarks, verifying the effectiveness of the proposed approach.

## [Factorized-FL: Personalized Federated Learning with Parameter Factorization & Similarity Matching](#)

- Wonyong Jeong Â· Sung Ju Hwang
- abstract@[open-review](#): In real-world federated learning scenarios, participants could have their own personalized labels which are incompatible with those from other clients, due to using different label permutations or tackling completely different tasks or domains. However, most existing FL approaches cannot effectively tackle such extremely heterogeneous scenarios since they often assume that (1) all participants use a synchronized set of labels, and (2) they train on the same tasks from the same domain. In this work, to tackle these challenges, we introduce Factorized-FL, which allows to effectively tackle label- and task-heterogeneous federated learning settings by factorizing the model parameters into a pair of rank-1 vectors, where one captures the common knowledge across different labels and tasks and the other captures knowledge specific to the task for each local model. Moreover, based on the distance in the client-specific vector space, Factorized-FL performs selective aggregation scheme to utilize only the knowledge from the relevant participants for each client. We extensively validate our method on both label- and domain-heterogeneous settings, on which it outperforms the state-of-the-art personalized federated learning methods.

## [Globally Convergent Policy Search for Output Estimation](#)

- Jack Umenberger Â· Max Simchowitz Â· Juan Perdomo Â· Kaiqing Zhang Â· Russ Tedrake
- abstract@[open-review](#): We introduce the first direct policy search algorithm which provably converges to the globally optimal dynamic filter for the classical problem of predicting the outputs of a linear dynamical system, given noisy, partial observations. Despite the ubiquity of partial observability in practice, theoretical guarantees for direct policy search algorithms, one of the backbones of modern reinforcement learning, have proven difficult to achieve. This is primarily due to the degeneracies which arise when optimizing over filters that maintain an internal state. In this paper, we provide a new perspective on this challenging problem based on the notion of informativity, which intuitively requires that all components of a filter's internal state are representative of the true state of the underlying dynamical system. We show that informativity overcomes the aforementioned degeneracy. Specifically, we propose a regularizer which explicitly enforces informativity, and establish that gradient descent on this regularized objective - combined with a "reconditioning step" converges to the globally optimal cost at a  $O(1/T)$  rate.

## [Bayesian Clustering of Neural Spiking Activity Using a Mixture of Dynamic Poisson Factor Analyzers](#)

- Ganchao Wei Â· Ian H Stevenson Â· Xiaojing Wang
- abstract@[open-review](#): Modern neural recording techniques allow neuroscientists to observe the spiking activity of many neurons simultaneously. Although previous work has illustrated how activity within and between known populations of neurons can be summarized by low-dimensional latent vectors, in many cases what determines a unique population may be unclear. Neurons differ in their anatomical location, but also, in their cell types and response properties. Moreover, multiple distinct populations may not be well described by a single low-dimensional, linear representation. To tackle these challenges, we develop a clustering method based on a mixture of dynamic Poisson factor analyzers (DPFA) model, with the number of clusters treated as an unknown parameter. To do the analysis of DPFA model, we propose a novel Markov chain Monte Carlo (MCMC) algorithm to efficiently sample its posterior distribution. Validating our proposed MCMC algorithm with simulations, we find that it can accurately recover the true clustering and latent states and is insensitive to the initial cluster assignments. We then apply the proposed mixture of DPFA model to multi-region experimental recordings, where we find that the proposed method can identify novel, reliable clusters of neurons based on their activity, and may, thus, be a useful tool for neural data analysis.

## [Provable Benefit of Multitask Representation Learning in Reinforcement Learning](#)

- Yuan Cheng Â· Songtao Feng Â· Jing Yang Â· Hong Zhang Â· Yingbin Liang
- abstract@[open-review](#): As representation learning becomes a powerful technique to reduce sample complexity in reinforcement learning (RL) in practice, theoretical understanding of its advantage is still limited. In this paper, we theoretically characterize the benefit of representation learning under the low-rank Markov decision process (MDP) model. We first study multitask low-rank RL (as upstream training), where all tasks share a common representation, and propose a new multitask reward-free algorithm called REFUEL. REFUEL learns both the transition kernel and the near-optimal policy for each task, and outputs a well-learned representation for downstream tasks. Our result demonstrates that multitask representation learning is provably more sample-efficient than learning each task individually, as long as the total number of tasks is above a certain threshold. We then study the downstream offline RL, where the agent is given a new task sharing the same representation as the upstream tasks and an offline dataset, and aims to find a near-optimal policy. We develop a sample-efficient algorithm with the suboptimality gap bounded by the estimation error of the learned representation in the upstream plus a vanishing term that decreases as the number of offline samples becomes large. Our result further captures the benefit of employing the learned representation from upstream training as opposed to learning the representation of the low-rank model directly. To the best of our knowledge, this is the first theoretical study that characterizes the benefit of representation learning in exploration-based reward-free multitask RL.

## [QUARK: Controllable Text Generation with Reinforced Unlearning](#)

- Ximing Lu Â· Sean Welleck Â· Liwei Jiang Â· Jack Hessel Â· Lianhui Qin Â· Peter West Â· Prithviraj Ammanabrolu Â· Yejin Choi
- abstract@[open-review](#): Large-scale language models often learn behaviors that are misaligned with user expectations. Generated text may contain offensive or toxic language, contain significant repetition, or be of a different sentiment than desired by the user. We consider the task of unlearning these misalignments by fine-tuning the language model on signals of what not to do. We introduce Quantized Reward Konditioning (Quark), an algorithm for optimizing a reward function that quantifies an (un)wanted property, while not straying too far from the original model. Quark alternates between (i) collecting samples with the current language model, (ii) sorting them into quantiles based on reward, with each quantile identified by a reward token

prepended to the language model's input, and (iii) using a standard language modeling loss on samples from each quantile conditioned on its reward token, while remaining nearby the original language model via a KL-divergence penalty. By conditioning on a high-reward token at generation time, the model generates text that exhibits less of the unwanted property. For unlearning toxicity, negative sentiment, and repetition, our experiments show that Quark outperforms both strong baselines and state-of-the-art reinforcement learning methods like PPO, while relying only on standard language modeling primitives.

## [ReCo: Retrieve and Co-segment for Zero-shot Transfer](#)

- Gyungin Shin · Weidi Xie · Samuel Albanie
- abstract@[open-review](#): Semantic segmentation has a broad range of applications, but its real-world impact has been significantly limited by the prohibitive annotation costs necessary to enable deployment. Segmentation methods that forgo supervision can side-step these costs, but exhibit the inconvenient requirement to provide labelled examples from the target distribution to assign concept names to predictions. An alternative line of work in language-image pre-training has recently demonstrated the potential to produce models that can both assign names across large vocabularies of concepts and enable zero-shot transfer for classification, but do not demonstrate commensurate segmentation abilities. In this work, we strive to achieve a synthesis of these two approaches that combines their strengths. We leverage the retrieval abilities of one such language-image pre-trained model, CLIP, to dynamically curate training sets from unlabelled images for arbitrary collections of concept names, and leverage the robust correspondences offered by modern image representations to co-segment entities among the resulting collections. The synthetic segment collections are then employed to construct a segmentation model (without requiring pixel labels) whose knowledge of concepts is inherited from the scalable pre-training process of CLIP. We demonstrate that our approach, termed Retrieve and Co-segment (ReCo) performs favourably to unsupervised segmentation approaches while inheriting the convenience of nameable predictions and zero-shot transfer. We also demonstrate ReCo's ability to generate specialist segmenters for extremely rare objects.

## [Learn to Match with No Regret: Reinforcement Learning in Markov Matching Markets](#)

- Yifei Min · Tianhao Wang · Ruitu Xu · Zhaoran Wang · Michael Jordan · Zhuoran Yang
- abstract@[open-review](#): We study a Markov matching market involving a planner and a set of strategic agents on the two sides of the market. At each step, the agents are presented with a dynamical context, where the contexts determine the utilities. The planner controls the transition of the contexts to maximize the cumulative social welfare, while the agents aim to find a myopic stable matching at each step. Such a setting captures a range of applications including ridesharing platforms. We formalize the problem by proposing a reinforcement learning framework that integrates optimistic value iteration with maximum weight matching. The proposed algorithm addresses the coupled challenges of sequential exploration, matching stability, and function approximation. We prove that the algorithm achieves sublinear regret.

## [Proximal Learning With Opponent-Learning Awareness](#)

- Stephen Zhao · Chris Lu · Roger Grosse · Jakob Foerster
- abstract@[open-review](#): Learning With Opponent-Learning Awareness (LOLA) (Foerster et al. [2018a]) is a multi-agent reinforcement learning algorithm that typically learns reciprocity-based cooperation in partially competitive environments. However, LOLA often fails to learn such behaviour on more complex policy spaces parameterized by neural networks, partly because the update rule is sensitive to the policy parameterization. This problem is especially pronounced in the opponent modeling setting, where the opponent's policy is unknown and must be inferred from observations; in such settings, LOLA is ill-specified because behaviorally equivalent opponent policies can result in non-equivalent updates. To address this shortcoming, we reinterpret LOLA as approximating a proximal operator, and then derive a new algorithm, Proximal LOLA (POLA), which uses the proximal formulation directly. Unlike LOLA, the POLA updates are parameterization invariant, in the sense that when the proximal objective has a unique optimum, behaviorally equivalent policies result in behaviorally equivalent updates. We then present practical approximations to the ideal POLA update. We evaluate these approximations in several partially competitive environments with function approximation and opponent modeling, empirically demonstrating POLA achieves reciprocity-based cooperation more reliably than LOLA.

## [FasterRisk: Fast and Accurate Interpretable Risk Scores](#)

- Jiachang Liu · Chudi Zhong · Boxuan Li · Margo Seltzer · Cynthia Rudin
- abstract@[open-review](#): Over the last century, risk scores have been the most popular form of predictive model used in healthcare and criminal justice. Risk scores are sparse linear models with integer coefficients; often these models can be memorized or placed on an index card. Typically, risk scores have been created either without data or by rounding logistic regression coefficients, but these methods do not reliably produce high-quality risk scores. Recent work used mathematical programming, which is computationally slow. We introduce an approach for efficiently producing a collection of high-quality risk scores learned from data. Our approach involves producing a pool of almost-optimal sparse continuous solutions, each with a different support set, using a beam-search algorithm. Each of these continuous solutions is transformed into a separate risk score through a "star search," where a range of multipliers are considered before rounding the coefficients sequentially to maintain low logistic loss. Our algorithm returns all of these high-quality risk scores for the user to consider. This method completes within minutes and can be impactful in a broad variety of applications.

## [Taming Fat-Tailed \("Heavier-Tailed" with Potentially Infinite Variance\) Noise in Federated Learning](#)

- Haibo Yang · Peiwen Qiu · Jia Liu
- abstract@[open-review](#): In recent years, federated learning (FL) has emerged as an important distributed machine learning paradigm to collaboratively learn a global model with multiple clients, while keeping data local and private. However, a key assumption in most existing works on FL algorithms' convergence analysis is that the noise in stochastic first-order information has a finite variance. Although this assumption covers all light-tailed (i.e., sub-exponential) and some heavy-tailed noise distributions (e.g., log-normal, Weibull, and some Pareto distributions), it fails for many fat-tailed noise distributions (i.e., ``heavier-tailed'' with potentially infinite variance) that have been empirically observed in the FL literature. To date, it remains unclear whether one can design convergent algorithms for FL systems that experience fat-tailed noise. This motivates us to fill this gap in this paper by proposing an algorithmic framework called  $\mathcal{FAT}$ - $\mathcal{Clipping}$  federated averaging with two-sided learning rates and  $\mathcal{clipping}$ , which contains two variants:  $\mathcal{FAT}$ - $\mathcal{Clipping}$  per-round ( $\mathcal{FAT}$ - $\mathcal{Clipping}$ ) and  $\mathcal{FAT}$ - $\mathcal{Clipping}$  per-iteration ( $\mathcal{FAT}$ - $\mathcal{Clipping}$ ). Specifically, for the largest  $\alpha \in (1, 2]$  such that the fat-tailed noise in FL still has a bounded  $\alpha$ -moment, we show that both variants achieve  $O((mT)^{\frac{1}{\alpha}})$  convergence rates in the strongly-convex and general non-convex settings, respectively, where  $m$  and  $T$  are the numbers of clients and communication rounds. Moreover, at the expense of more clipping operations compared to  $\mathcal{FAT}$ - $\mathcal{PR}$ ,  $\mathcal{FAT}$ - $\mathcal{Clipping}$  further enjoys a linear speedup effect with respect to the number of local updates at each client and being lower-bound-matching (i.e., order-optimal). Collectively, our results advance the understanding of designing efficient algorithms for FL systems that exhibit fat-tailed first-order oracle information.

## [Deep Architecture Connectivity Matters for Its Convergence: A Fine-Grained Analysis](#)

- Wuyang Chen · Wei Huang · Xinyu Gong · Boris Hanin · Zhangyang Wang
- abstract@[open-review](#): Advanced deep neural networks (DNNs), designed by either human or AutoML algorithms, are growing increasingly complex. Diverse operations are connected by complicated connectivity patterns, e.g., various types of skip connections. Those topological compositions are

empirically effective and observed to smooth the loss landscape and facilitate the gradient flow in general. However, it remains elusive to derive any principled understanding of their effects on the DNN capacity or trainability, and to understand why or in which aspect one specific connectivity pattern is better than another. In this work, we theoretically characterize the impact of connectivity patterns on the convergence of DNNs under gradient descent training in fine granularity. By analyzing a wide network's Neural Network Gaussian Process (NNGP), we are able to depict how the spectrum of an NNGP kernel propagates through a particular connectivity pattern, and how that affects the bound of convergence rates. As one practical implication of our results, we show that by a simple filtration of "unpromising" connectivity patterns, we can trim down the number of models to evaluate, and significantly accelerate the large-scale neural architecture search without any overhead.

## [A Simple and Provably Efficient Algorithm for Asynchronous Federated Contextual Linear Bandits](#)

- Jiafan He · Tianhao Wang · Yifei Min · Quanquan Gu
- abstract@[open-review](#): We study federated contextual linear bandits, where  $M$  agents cooperate with each other to solve a global contextual linear bandit problem with the help of a central server. We consider the asynchronous setting, where all agents work independently and the communication between one agent and the server will not trigger other agents' communication. We propose a simple algorithm named FedLinUCB based on the principle of optimism. We prove that the regret of FedLinUCB is bounded by  $\tilde{O}(d\sqrt{\sum_{m=1}^M T_m})$  and the communication complexity is  $\tilde{O}(dM^2)$ , where  $d$  is the dimension of the contextual vector and  $T_m$  is the total number of interactions with the environment by agent  $m$ . To the best of our knowledge, this is the first provably efficient algorithm that allows fully asynchronous communication for federated linear bandits, while achieving the same regret guarantee as in the single-agent setting.

## [APG: Adaptive Parameter Generation Network for Click-Through Rate Prediction](#)

- Bencheng Yan · Pengjie Wang · Kai Zhang · Feng Li · Hongbo Deng · Jian Xu · Bo Zheng
- abstract@[open-review](#): In many web applications, deep learning-based CTR prediction models (deep CTR models for short) are widely adopted. Traditional deep CTR models learn patterns in a static manner, i.e., the network parameters are the same across all the instances. However, such a manner can hardly characterize each of the instances which may have different underlying distributions. It actually limits the representation power of deep CTR models, leading to sub-optimal results. In this paper, we propose an efficient, effective, and universal module, named as Adaptive Parameter Generation network (APG), which can dynamically generate parameters for deep CTR models on-the-fly based on different instances. Extensive experimental evaluation results show that APG can be applied to a variety of deep CTR models and significantly improve their performance. Meanwhile, APG can reduce the time cost by 38.7% and memory usage by 96.6% compared to a regular deep CTR model. We have deployed APG in the industrial sponsored search system and achieved 3% CTR gain and 1% RPM gain respectively.

## [Differentially Private Learning with Margin Guarantees](#)

- Raef Bassily · Mehryar Mohri · Ananda Theertha Suresh
- abstract@[open-review](#): We present a series of new differentially private (DP) algorithms with dimension-independent margin guarantees. For the family of linear hypotheses, we give a pure DP learning algorithm that benefits from relative deviation margin guarantees, as well as an efficient DP learning algorithm with margin guarantees. We also present a new efficient DP learning algorithm with margin guarantees for kernel-based hypotheses with shift-invariant kernels, such as Gaussian kernels, and point out how our results can be extended to other kernels using oblivious sketching techniques. We further give a pure DP learning algorithm for a family of feed-forward neural networks for which we prove margin guarantees that are independent of the input dimension. Additionally, we describe a general label DP learning algorithm, which benefits from relative deviation margin bounds and is applicable to a broad family of hypothesis sets, including that of neural networks. Finally, we show how our DP learning algorithms can be augmented in a general way to include model selection, to select the best confidence margin parameter.

## [Oracle-Efficient Online Learning for Smoothed Adversaries](#)

- Nika Haghtalab · Yanjun Han · Abhishek Shetty · Kunhe Yang
- abstract@[open-review](#): In this paper, we study oracle-efficient algorithms for smoothed analysis of online learning. In this setting, an adversary is constrained to generating samples from distributions whose density is upper bounded by  $1/\sigma$  times the uniform density. Given access to an offline optimization oracle, we give the first computationally efficient online algorithms whose sublinear regret depends only on the pseudo/VC dimension  $d$  of the class and the smoothness parameter  $\sigma$ . In particular, we achieve oracle-efficient regret bounds of  $O(\sqrt{T\sigma^{-1}})$  for learning real-valued functions and  $O(\sqrt{T\sigma^{-1/2}})$  for learning binary-valued functions. This contrasts the computational separation between online learning with worst-case adversaries and offline learning established by [HK16]. In the binary setting, our algorithms also achieve improved bounds for worst-case setting with small domains. In particular, we give an oracle-efficient algorithm with regret of  $O(\sqrt{T\mathcal{X}^{1/2}})$ , which is a refinement of the earlier  $O(\sqrt{T\mathcal{X}})$  bound by [DS16].

## [Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering](#)

- Pan Lu · Swaroop Mishra · Tanglin Xia · Liang Qiu · Kai-Wei Chang · Song-Chun Zhu · Oyvind Tafjord · Peter Clark · Ashwin Kalyan
- abstract@[open-review](#): When answering a question, humans utilize the information available across different modalities to synthesize a consistent and complete chain of thought (CoT). This process is normally a black box in the case of deep learning models like large-scale language models. Recently, science question benchmarks have been used to diagnose the multi-hop reasoning ability and interpretability of an AI system. However, existing datasets fail to provide annotations for the answers, or are restricted to the textual-only modality, small scales, and limited domain diversity. To this end, we present Science Question Answering (SQA), a new benchmark that consists of ~21k multimodal multiple choice questions with a diverse set of science topics and annotations of their answers with corresponding lectures and explanations. We further design language models to learn to generate lectures and explanations as the chain of thought (CoT) to mimic the multi-hop reasoning process when answering SQA questions. SQA demonstrates the utility of CoT in language models, as CoT improves the question answering performance by 1.20% in few-shot GPT-3 and 3.99% in fine-tuned UnifiedQA. We also explore the upper bound for models to leverage explanations by feeding those in the input; we observe that it improves the few-shot performance of GPT-3 by 18.96%. Our analysis further shows that language models, similar to humans, benefit from explanations to learn from fewer data and achieve the same performance with just 40% of the data.

## [Foundation Posteriors for Approximate Probabilistic Inference](#)

- Mike Wu · Noah Goodman
- abstract@[open-review](#): Probabilistic programs provide an expressive representation language for generative models. Given a probabilistic program, we are interested in the task of posterior inference: estimating a latent variable given a set of observed variables. Existing techniques for inference in probabilistic programs often require choosing many hyper-parameters, are computationally expensive, and/or only work for restricted classes of programs. Here we formulate inference as masked language modeling: given a program, we generate a supervised dataset of variables and assignments, and randomly mask a subset of the assignments. We then train a neural network to unmask the random values, defining an approximate posterior distribution. By optimizing a single neural network across a range of programs we amortize the cost of training, yielding a "foundation" posterior able to do zero-shot inference for new programs. The foundation posterior can also be fine-tuned for a particular program and dataset by optimizing a variational inference objective. We show the efficacy of the approach, zero-shot and fine-tuned, on a benchmark of STAN programs.

## [Anchor-Changing Regularized Natural Policy Gradient for Multi-Objective Reinforcement Learning](#)

- Ruida Zhou · Tao Liu · Dileep Kalathil · P. R. Kumar · Chao Tian
- abstract@[open-review](#): We study policy optimization for Markov decision processes (MDPs) with multiple reward value functions, which are to be jointly optimized according to given criteria such as proportional fairness (smooth concave scalarization), hard constraints (constrained MDP), and max-min trade-off. We propose an Anchor-changing Regularized Natural Policy Gradient (ARNPG) framework, which can systematically incorporate ideas from well-performing first-order methods into the design of policy optimization algorithms for multi-objective MDP problems. Theoretically, the designed algorithms based on the ARNPG framework achieve  $\tilde{O}(1/T)$  global convergence with exact gradients. Empirically, the ARNPG-guided algorithms also demonstrate superior performance compared to some existing policy gradient-based approaches in both exact gradients and sample-based scenarios.

## [A Best-of-Both-Worlds Algorithm for Bandits with Delayed Feedback](#)

- Saeed Masoudian · Julian Zimmert · Yevgeny Seldin
- abstract@[open-review](#): We present a modified tuning of the algorithm of Zimmert and Seldin [2020] for adversarial multiarmed bandits with delayed feedback, which in addition to the minimax optimal adversarial regret guarantee shown by Zimmert and Seldin [2020] simultaneously achieves a near-optimal regret guarantee in the stochastic setting with fixed delays. Specifically, the adversarial regret guarantee is  $\mathcal{O}(\sqrt{TK} + \sqrt{dT \log K})$ , where  $T$  is the time horizon,  $K$  is the number of arms, and  $d$  is the fixed delay, whereas the stochastic regret guarantee is  $\mathcal{O}(\sum_{i \neq i^*} \frac{1}{\Delta_i} \log(T) + \frac{d}{\Delta_{i^*}} + dK^{1/3} \log K)$ , where  $\Delta_i$  are the suboptimality gaps. We also present an extension of the algorithm to the case of arbitrary delays, which is based on an oracle knowledge of the maximal delay  $d_{\max}$  and achieves  $\mathcal{O}(\sqrt{TK} + \sqrt{D \log K} + d_{\max} K^{1/3} \log K)$  regret in the adversarial regime, where  $D$  is the total delay, and  $\mathcal{O}(\sum_{i \neq i^*} \frac{1}{\Delta_i} \log(T) + \frac{\sigma_{\max}}{\Delta_{i^*}} + d_{\max} K^{1/3} \log K)$  regret in the stochastic regime, where  $\sigma_{\max}$  is the maximal number of outstanding observations. Finally, we present a lower bound that matches regret upper bound achieved by the skipping technique of Zimmert and Seldin [2020] in the adversarial setting.

## [Pre-Trained Language Models for Interactive Decision-Making](#)

- Shuang Li · Xavier Puig · Chris Paxton · Yilun Du · Clinton Wang · Linxi Fan · Tao Chen · De-An Huang · Ekin Akyurek · Anima Anandkumar · Jacob Andreas · Igor Mordatch · Antonio Torralba · Yuke Zhu
- abstract@[open-review](#): Language model (LM) pre-training is useful in many language processing tasks. But can pre-trained LMs be further leveraged for more general machine learning problems? We propose an approach for using LMs to scaffold learning and generalization in general sequential decision-making problems. In this approach, goals and observations are represented as a sequence of embeddings, and a policy network initialized with a pre-trained LM predicts the next action. We demonstrate that this framework enables effective combinatorial generalization across different environments and supervisory modalities. We begin by assuming access to a set of expert demonstrations, and show that initializing policies with LMs and fine-tuning them via behavior cloning improves task completion rates by 43.6% in the VirtualHome environment. We then examine how our framework may be used in environments without pre-collected expert data. To do this, we integrate an active data gathering procedure into pre-trained LMs. The agent iteratively learns by interacting with the environment, relabeling the language goal of past "failed" experiences, and updating the policy in a self-supervised loop. The active data gathering procedure also enables effective combinatorial generalization, outperforming the best baseline by 25.1%. Finally, we explain these results by investigating three possible factors underlying the effectiveness of the LM-based policy. We find that sequential input representations (vs. fixed-dimensional feature vectors) and favorable weight initialization are both important for generalization. Surprisingly, however, the format of the policy inputs encoding (e.g. as a natural language string vs. an arbitrary sequential encoding) has little influence. Together, these results suggest that language modeling induces representations that are useful for modeling not just language, but also goals and plans; these representations can aid learning and generalization even outside of language processing.

## [VoiceBox: Privacy through Real-Time Adversarial Attacks with Audio-to-Audio Models](#)

- Patrick O'Reilly · Andreas Bugler · Keshav Bhandari · Max Morrison · Bryan Pardo
- abstract@[open-review](#): As governments and corporations adopt deep learning systems to collect and analyze user-generated audio data, concerns about security and privacy naturally emerge in areas such as automatic speaker recognition. While audio adversarial examples offer one route to mislead or evade these invasive systems, they are typically crafted through time-intensive offline optimization, limiting their usefulness in streaming contexts. Inspired by architectures for audio-to-audio tasks such as denoising and speech enhancement, we propose a neural network model capable of adversarially modifying a user's audio stream in real-time. Our model learns to apply a time-varying finite impulse response (FIR) filter to outgoing audio, allowing for effective and inconspicuous perturbations on a small fixed delay suitable for streaming tasks. We demonstrate our model is highly effective at de-identifying user speech from speaker recognition and able to transfer to an unseen recognition system. We conduct a perceptual study and find that our method produces perturbations significantly less perceptible than baseline anonymization methods, when controlling for effectiveness. Finally, we provide an implementation of our model capable of running in real-time on a single CPU thread. Audio examples and code can be found at <https://master.d3hvbnf7qxjtf.amplifyapp.com/>.

## [Micro and Macro Level Graph Modeling for Graph Variational Auto-Encoders](#)

- Kiarash Zahirnia · Parmis Naddaf · Oliver Schulte · Ke Li
- abstract@[open-review](#): Generative models for graph data are an important research topic in machine learning. Graph data comprise two levels that are typically analyzed separately: node-level properties such as the existence of a link between a pair of nodes, and global aggregate graph-level statistics, such as motif counts. This paper proposes a new multi-level framework that jointly models node-level properties and graph-level statistics, as mutually reinforcing sources of information. We introduce a new micro-macro training objective for graph generation that combines node-level and graph-level losses. We utilize the micro-macro objective to improve graph generation with a GraphVAE [41], a well-established model based on graph-level latent variables, that provides fast training and generation time for medium-sized graphs. Our experiments show that adding micro-macro modeling to the GraphVAE model improves graph quality scores up to 2 orders of magnitude on five benchmark datasets, while maintaining the GraphVAE generation speed advantage.

## [LIFT: Language-Interfaced FineTuning for Non-language Machine Learning Tasks](#)

- Tuan Dinh · Yuchen Zeng · Ruisu Zhang · Ziqian Lin · Michael Gira · Shashank Rajput · Jy-yong Sohn · Dimitris Papailiopoulos · Kangwook Lee
- abstract@[open-review](#): Finetuning pretrained language models (LMs) has become a norm for learning various downstream tasks. While it is feasible to finetune LMs for language down-stream tasks without making any architectural changes, most existing finetuning approaches for non-language tasks rely on task-specific designs for input, output layers, and loss functions. A natural question arises – Can language model finetuning solve non-language downstream tasks without changing models' architecture or loss function? To answer this question, we study the efficacy and limitations of Language-Interfaced FineTuning (LIFT) for non-language tasks by conducting an extensive empirical study on a suite of non-language classification and regression tasks. LIFT does not make any changes to the model architecture or loss function, and it solely relies on the natural language interface, truly enabling "zero-code machine learning with language models". We find that LIFT performs relatively well across a wide range of low-dimensional classification and regression tasks, matching the performances of the best models in many cases, especially for the classification tasks. We thoroughly

study fundamental properties of LIFT, including the inductive bias, sample efficiency, ability to extrapolate, robustness to noises and corrupted labels, and adversarial robustness. We also analyze a few unique properties specific to LIFT – non-deterministic predictions and how to use them, and sample-efficient context-aware learning via appropriate prompting or two-stage finetuning. We provide discussions on limitations and open questions toward making LIFT more effective and efficient.

## [Momentum Aggregation for Private Non-convex ERM](#)

- Hoang Tran · Ashok Cutkosky
- abstract@[open-review](#): We introduce new algorithms and convergence guarantees for privacy-preserving non-convex Empirical Risk Minimization (ERM) on smooth  $d$ -dimensional objectives. We develop an improved sensitivity analysis of stochastic gradient descent on smooth objectives that exploits the recurrence of examples in different epochs. By combining this new approach with recent analysis of momentum with private aggregation techniques, we provide an  $(\epsilon, \delta)$ -differential private algorithm that finds a gradient of norm  $O(\left(\frac{d^{1/3}}{\epsilon}\right)^{N^{2/3}})$  gradient evaluations, improving the previous best gradient bound of  $\tilde{O}(\frac{d^{1/4}}{\epsilon} \sqrt{N})$ .

## [Data Augmentation MCMC for Bayesian Inference from Privatized Data](#)

- Nianqiao Ju · Jordan Awan · Ruobin Gong · Vinayak Rao
- abstract@[open-review](#): Differentially private mechanisms protect privacy by introducing additional randomness into the data. Restricting access to only the privatized data makes it challenging to perform valid statistical inference on parameters underlying the confidential data. Specifically, the likelihood function of the privatized data requires integrating over the large space of confidential databases and is typically intractable. For Bayesian analysis, this results in a posterior distribution that is doubly intractable, rendering traditional MCMC techniques inapplicable. We propose an MCMC framework to perform Bayesian inference from the privatized data, which is applicable to a wide range of statistical models and privacy mechanisms. Our MCMC algorithm augments the model parameters with the unobserved confidential data, and alternately updates each one conditional on the other. For the potentially challenging step of updating the confidential data, we propose a generic approach that exploits the privacy guarantee of the mechanism to ensure efficiency. In particular, we give results on the computational complexity, acceptance rate, and mixing properties of our MCMC. We illustrate the efficacy and applicability of our methods on a negative-Bayes log-linear model as well as on a linear regression model.

## [Triangulation candidates for Bayesian optimization](#)

- Robert Gramacy · Annie Sauer · Nathan Wycoff
- abstract@[open-review](#): Bayesian optimization involves "inner optimization" over a new-data acquisition criterion which is non-convex/highly multi-modal, may be non-differentiable, or may otherwise thwart local numerical optimizers. In such cases it is common to replace continuous search with a discrete one over random candidates. Here we propose using candidates based on a Delaunay triangulation of the existing input design. We detail the construction of these "tricands" and demonstrate empirically how they outperform both numerically optimized acquisitions and random candidate-based alternatives, and are well-suited for hybrid schemes, on benchmark synthetic and real simulation experiments.

## [Langevin Autoencoders for Learning Deep Latent Variable Models](#)

- Shohei Taniguchi · Yusuke Iwasawa · Wataru Kumagai · Yutaka Matsuo
- abstract@[open-review](#): Markov chain Monte Carlo (MCMC), such as Langevin dynamics, is valid for approximating intractable distributions. However, its usage is limited in the context of deep latent variable models owing to costly datapoint-wise sampling iterations and slow convergence. This paper proposes the amortized Langevin dynamics (ALD), wherein datapoint-wise MCMC iterations are entirely replaced with updates of an encoder that maps observations into latent variables. This amortization enables efficient posterior sampling without datapoint-wise iterations. Despite its efficiency, we prove that ALD is valid as a MCMC algorithm, whose Markov chain has the target posterior as a stationary distribution under mild assumptions. Based on the ALD, we also present a new deep latent variable model named the Langevin autoencoder (LAE). Interestingly, the LAE can be implemented by slightly modifying the traditional autoencoder. Using multiple synthetic datasets, we first validate that ALD can properly obtain samples from target posteriors. We also evaluate the LAE on the image generation task, and show that our LAE can outperform existing methods based on variational inference, such as the variational autoencoder, and other MCMC-based methods in terms of the test likelihood.

## [Object-Category Aware Reinforcement Learning](#)

- Qi Yi · Rui Zhang · shaohui peng · Jiaming Guo · Xing Hu · Zidong Du · xishan zhang · Qi Guo · Yunji Chen
- abstract@[open-review](#): Object-oriented reinforcement learning (OORL) is a promising way to improve the sample efficiency and generalization ability over standard RL. Recent works that try to solve OORL tasks without additional feature engineering mainly focus on learning the object representations and then solving tasks via reasoning based on these object representations. However, none of these works tries to explicitly model the inherent similarity between different object instances of the same category. Objects of the same category should share similar functionalities; therefore, the category is the most critical property of an object. Following this insight, we propose a novel framework named Object-Category Aware Reinforcement Learning (OCARL), which utilizes the category information of objects to facilitate both perception and reasoning. OCARL consists of three parts: (1) Category-Aware Unsupervised Object Discovery (UOD), which discovers the objects as well as their corresponding categories; (2) Object-Category Aware Perception, which encodes the category information and is also robust to the incompleteness of (1) at the same time; (3) Object-Centric Modular Reasoning, which adopts multiple independent and object-category-specific networks when reasoning based on objects. Our experiments show that OCARL can improve both the sample efficiency and generalization in the OORL domain.

## [Improving Zero-Shot Generalization in Offline Reinforcement Learning using Generalized Similarity Functions](#)

- Bogdan Mazoure · Ilya Kostrikov · Ofir Nachum · Jonathan Tompson
- abstract@[open-review](#): Reinforcement learning (RL) agents are widely used for solving complex sequential decision-making tasks, but still exhibit difficulty generalizing to scenarios not seen during training. While prior online approaches demonstrated that using additional signals beyond the reward function can lead to better generalization capabilities in RL agents, i.e. using self-supervised learning (SSL), they struggle in the offline RL setting, i.e. learning from a static dataset. We show that the performance of online algorithms for generalization in RL can be hindered in the offline setting due to poor estimation of similarity between observations. We propose a new theoretically-motivated framework called Generalized Similarity Functions (GSF), which uses contrastive learning to train an offline RL agent to aggregate observations based on the similarity of their expected future behavior, where we quantify this similarity using generalized value functions. We show that GSF is general enough to recover existing SSL objectives while improving zero-shot generalization performance on two complex pixel-based offline RL benchmarks.

## [On the Statistical Efficiency of Reward-Free Exploration in Non-Linear RL](#)

- Jinglin Chen · Aditya Modi · Akshay Krishnamurthy · Nan Jiang · Alekh Agarwal
- abstract@[open-review](#): We study reward-free reinforcement learning (RL) under general non-linear function approximation, and establish sample efficiency and hardness results under various standard structural assumptions. On the positive side, we propose the RFOLIVE (Reward-Free OLIVE)

algorithm for sample-efficient reward-free exploration under minimal structural assumptions, which covers the previously studied settings of linear MDPs (Jin et al., 2020b), linear completeness (Zanette et al., 2020b) and low-rank MDPs with unknown representation (Modi et al., 2021). Our analyses indicate that the explorability or reachability assumptions, previously made for the latter two settings, are not necessary statistically for reward-free exploration. On the negative side, we provide a statistical hardness result for both reward-free and reward-aware exploration under linear completeness assumptions when the underlying features are unknown, showing an exponential separation between low-rank and linear completeness settings.

## [Expectation-Maximization Contrastive Learning for Compact Video-and-Language Representations](#)

- Peng Jin · Fa Jin Huang · Fenglin Liu · Xian Wu · Shen Ge · Guoli Song · David Clifton · Jie Chen
- abstract@[open-review](#): Most video-and-language representation learning approaches employ contrastive learning, e.g., CLIP, to project the video and text features into a common latent space according to the semantic similarities of text-video pairs. However, such learned shared latent spaces are not often optimal, and the modality gap between visual and textual representation can not be fully eliminated. In this paper, we propose Expectation-Maximization Contrastive Learning (EMCL) to learn compact video-and-language representations. Specifically, we use the Expectation-Maximization algorithm to find a compact set of bases for the latent space, where the features could be concisely represented as the linear combinations of these bases. Such feature decomposition of video-and-language representations reduces the rank of the latent space, resulting in increased representing power for the semantics. Extensive experiments on three benchmark text-video retrieval datasets prove that our EMCL can learn more discriminative video-and-language representations than previous methods, and significantly outperform previous state-of-the-art methods across all metrics. More encouragingly, the proposed method can be applied to boost the performance of existing approaches either as a jointly training layer or an out-of-the-box inference module with no extra training, making it easy to be incorporated into any existing methods.

## [Memory safe computations with XLA compiler](#)

- Artem Artemev · Tilman Roeder · Mark van der Wilk
- abstract@[open-review](#): Software packages like TensorFlow and Pytorch are designed to support linear algebra operations, and their speed and usability determine their success. However, by prioritising speed, they often neglect memory requirements. As a consequence, the implementations of memory-intensive algorithms that are convenient in terms of software design can often not be run for large problems due to memory overflows. Memory-efficient solutions require complex programming approaches with significant logic outside the computational framework. This impairs the adoption and use of such algorithms. To address this, we developed an XLA compiler extension that adjusts the computational data-flow representation of an algorithm according to a user-specified memory limit. We show that k-nearest neighbour (kNN) and sparse Gaussian process regression (SGPR) methods can be run at a much larger scale on a single device, where standard implementations would have failed. Our approach leads to better use of hardware resources. We believe that further focus on removing memory constraints at a compiler level will widen the range of machine learning methods that can be developed in the future.

## [Parameter-Efficient Image-to-Video Transfer Learning](#)

- Junting Pan · Ziyi Lin · Xiatian Zhu · Jing Shao · Hongsheng Li
- abstract@[open-review](#): Capitalizing on large pre-trained models for various downstream tasks of interest have recently emerged with promising performance. Due to the ever-growing model size, the standard full fine-tuning based task adaptation strategy becomes prohibitively costly in terms of model training and storage. This has led to a new research direction in parameter-efficient transfer learning. However, existing attempts typically focus on downstream tasks from the same modality (e.g., image understanding) of the pre-trained model. This creates a limit because in some specific modalities, (e.g., video understanding) such a strong pre-trained model with sufficient knowledge is less or not available. In this work, we investigate such a novel cross-modality transfer learning setting, namely parameter-efficient image-to-video transfer learning. To solve this problem, we propose a new Spatio-Temporal Adapter (ST-Adapter) for parameter-efficient fine-tuning per video task. With a built-in spatio-temporal reasoning capability in a compact design, ST-Adapter enables a pre-trained image model without temporal knowledge to reason about dynamic video content at a small ~8% per-task parameter cost, requiring approximately 20 times fewer updated parameters compared to previous work. Extensive experiments on video action recognition tasks show that our ST-Adapter can match or even outperform the strong full fine-tuning strategy and state-of-the-art video models, whilst enjoying the advantage of parameter efficiency.

## [Renyi Differential Privacy of Propose-Test-Release and Applications to Private and Robust Machine Learning](#)

- Jiachen T. Wang · Saeed Mahloujifar · Shouda Wang · Ruoxi Jia · Prateek Mittal
- abstract@[open-review](#): Propose-Test-Release (PTR) is a differential privacy framework that works with local sensitivity of functions, instead of their global sensitivity. This framework is typically used for releasing robust statistics such as median or trimmed mean in a differentially private manner. While PTR is a common framework introduced over a decade ago, using it in applications such as robust SGD where we need many adaptive robust queries is challenging. This is mainly due to the lack of \Renyi Differential Privacy (RDP) analysis, an essential ingredient underlying the moments accountant approach for differentially private deep learning. In this work, we generalize the standard PTR and derive the first RDP bound for it. We show that our RDP bound for PTR yields tighter DP guarantees than the directly analyzed \$(\varepsilon, \delta)\$-DP. We also derive the algorithm-specific privacy amplification bound of PTR under subsampling. We show that our bound is much tighter than the general upper bound and close to the lower bound. Our RDP bounds enable tighter privacy loss calculation for the composition of many adaptive runs of PTR. As an application of our analysis, we show that PTR and our theoretical results can be used to design differentially private variants for byzantine robust training algorithms that use robust statistics for gradients aggregation. We conduct experiments on the settings of label, feature, and gradient corruption across different datasets and architectures. We show that PTR-based private and robust training algorithm significantly improves the utility compared with the baseline.

## [Spherical Channels for Modeling Atomic Interactions](#)

- Larry Zitnick · Abhishek Das · Adeesh Kolluru · Janice Lan · Muhammed Shuaibi · Anuroop Sriram · Zachary Ulissi · Brandon Wood
- abstract@[open-review](#): Modeling the energy and forces of atomic systems is a fundamental problem in computational chemistry with the potential to help address many of the world's most pressing problems, including those related to energy scarcity and climate change. These calculations are traditionally performed using Density Functional Theory, which is computationally very expensive. Machine learning has the potential to dramatically improve the efficiency of these calculations from days or hours to seconds. We propose the Spherical Channel Network (SCN) to model atomic energies and forces. The SCN is a graph neural network where nodes represent atoms and edges their neighboring atoms. The atom embeddings are a set of spherical functions, called spherical channels, represented using spherical harmonics. We demonstrate, that by rotating the embeddings based on the 3D edge orientation, more information may be utilized while maintaining the rotational equivariance of the messages. While equivariance is a desirable property, we find that by relaxing this constraint in both message passing and aggregation, improved accuracy may be achieved. We demonstrate state-of-the-art results on the large-scale Open Catalyst dataset in both energy and force prediction for numerous tasks and metrics.

## [House of Cans: Covert Transmission of Internal Datasets via Capacity-Aware Neuron Steganography](#)

- Xudong Pan · Shengyao Zhang · Mi Zhang · Yifan Yan · Min Yang
- abstract@[open-review](#): In this paper, we present a capacity-aware neuron steganography scheme (i.e., Cans) to covertly transmit multiple private machine learning (ML) datasets via a scheduled-to-publish deep neural network (DNN) as \textit{the carrier model}. Unlike existing steganography schemes which treat the DNN parameters as bit strings, Cans for the first time exploits the learning capacity of the carrier model via a novel parameter sharing

mechanism. Extensive evaluation shows, Cans is the first working scheme which can covertly transmit over \$10000\$ real-world data samples within a carrier model which has \$100\times\$ less parameters than the total size of the stolen data, and simultaneously transmit multiple heterogeneous datasets spanning visual, text and audio applications within a single carrier model, under a trivial distortion rate (\$<10^{-5}\$) and with almost no utility loss on the carrier model (\$<1\%\$). Besides, Cans implements by-design redundancy to be resilient against common post-processing techniques on the carrier model before the publishing.

## [Proviable Defense against Backdoor Policies in Reinforcement Learning](#)

- Shubham Bharti · Xuezhou Zhang · Adish Singla · Jerry Zhu
- abstract@[open-review](#): We propose a provable defense mechanism against backdoor policies in reinforcement learning. A backdoor policy is a security threat where an adversary publishes a seemingly well-behaved policy which in fact allows hidden triggers. During deployment, the adversary can modify observed states in a particular way to trigger unexpected actions and harm the agent. We assume the agent does not have the resources to re-train a good policy. Instead, our defense mechanism sanitizes the backdoor policy by projecting observed states to a safe subspace, estimated from a small number of interactions with a clean (non-triggered) environment. Our sanitized policy achieves  $\epsilon$ -optimality in the face of triggers in any and all rounds, provided the number of clean interactions is  $O(\frac{D}{(1-\gamma)^4 \epsilon^2})$  where  $\gamma$  is the discounting factor and  $D$  is the dimension of state space. Empirically, we show that our sanitization defense performs well on two Atari games.

## [Generalised Implicit Neural Representations](#)

- Daniele Grattarola · Pierre Vandergheynst
- abstract@[open-review](#): We consider the problem of learning implicit neural representations (INRs) for signals on non-Euclidean domains. In the Euclidean case, INRs are trained on a discrete sampling of a signal over a regular lattice. Here, we assume that the continuous signal exists on some unknown topological space from which we sample a discrete graph. In the absence of a coordinate system to identify the sampled nodes, we propose approximating their location with a spectral embedding of the graph. This allows us to train INRs without knowing the underlying continuous domain, which is the case for most graph signals in nature, while also making the INRs independent of any choice of coordinate system. We show experiments with our method on various real-world signals on non-Euclidean domains.

## [Riemannian Score-Based Generative Modelling](#)

- Valentin De Bortoli · Emile Mathieu · Michael Hutchinson · James Thornton · Yee Whye Teh · Arnaud Doucet
- abstract@[open-review](#): Score-based generative models (SGMs) are a powerful class of generative models that exhibit remarkable empirical performance. Score-based generative modelling (SGM) consists of a "noising" stage, whereby a diffusion is used to gradually add Gaussian noise to data, and a generative model, which entails a "denoising" process defined by approximating the time-reversal of the diffusion. Existing SGMs assume that data is supported on a Euclidean space, i.e. a manifold with flat geometry. In many domains such as robotics, geoscience or protein modelling, data is often naturally described by distributions living on Riemannian manifolds and current SGM techniques are not appropriate. We introduce here Riemannian Score-based Generative Models (RSGMs), a class of generative models extending SGMs to Riemannian manifolds. We demonstrate our approach on a variety of compact manifolds, and in particular with earth and climate science spherical data.

## [Multitasking Models are Robust to Structural Failure: A Neural Model for Bilingual Cognitive Reserve](#)

- Giannis Daras · Negin Raoof · Zoi Gkalitsiou · Alex Dimakis
- abstract@[open-review](#): We find a surprising connection between multitask learning and robustness to neuron failures. Our experiments show that bilingual language models retain higher performance under various neuron perturbations, such as random deletions, magnitude pruning and weight noise. Our study is motivated by research in cognitive science showing that symptoms of dementia and cognitive decline appear later in bilingual speakers compared to monolingual patients with similar brain damage, a phenomenon called bilingual cognitive reserve. Our language model experiments replicate this phenomenon on bilingual GPT-2 and other models. We provide a theoretical justification of this robustness by mathematically analyzing linear representation learning and showing that multitasking creates more robust representations.

## [Bringing Efficiency and Interpretability to Learned TCP Congestion Control](#)

- S P Sharan · Wenqing Zheng · Kuo-Feng Hsu · Jiarong Xing · Ang Chen · Zhangyang Wang
- abstract@[open-review](#): Recent research in TCP congestion control (CC) has witnessed tremendous success with deep reinforcement learning (RL) approaches, which use feedforward neural networks (NN) to tackle complex environment conditions and make better decisions. However, these "black box" policies lack interpretability, and reliability and, more importantly, cannot operate under the TCP datapath's ultra-contingent latency and computational constraints. This paper proposes a novel two-stage solution to achieve the best of both worlds: first to train a deep RL agent, then distill its (over-)parameterized NN policy into white-box, light-weight rules in the form of symbolic expressions that are much easier to understand and to implement in constrained environments. At the core of our proposal is a novel symbolic branching algorithm that allows the rule to be context-aware of various network conditions, eventually converting the NN policy into a symbolic tree. The distilled symbolic rules preserve and often improve performance over state-of-the-art NN policies while being orders of magnitude faster and interpretable. We validate the performance of our distilled symbolic rules on both simulation and emulation network systems. Our code will be released upon acceptance.

## [Museformer: Transformer with Fine- and Coarse-Grained Attention for Music Generation](#)

- Botao Yu · Peiling Lu · Rui Wang · Wei Hu · Xu Tan · Wei Ye · Shikun Zhang · Tao Qin · Tie-Yan Liu
- abstract@[open-review](#): Music sequences are typically very long, e.g., with over 10,000 tokens, making them hard to model by Transformer due to its quadratic complexity of self-attention. Although there are many Transformer variants aiming at modeling long sequences in natural language processing, directly using them in music generation is suboptimal, because music has both short-term structures and long-term structures, which is very different from text. In this paper, we propose Museformer, a Transformer with a novel fine- and coarse-grained attention for symbolic music generation. Specifically, with fine-grained attention, a token of a specific bar directly attends to all the tokens of the bars that are most relevant in regard to music structures (e.g., the previous 1st, 2nd, 4th and 8th bars, selected based on musical knowledge); with coarse-grained attention, a token only attends to the summarization of the other bars rather than each token of them so as to reduce computational cost. The advantages of our model are two-fold. First, it can capture both music structure-related correlations via the fine-grained attention, and other contextual information via the coarse-grained attention. Second, it is efficient and can well model over 4X longer music sequences compared to its full-attention counterpart. Both objective and subjective experimental results demonstrate its ability to generate long music sequences with high quality and better structures. Our generated music samples can be found at <https://museformer.github.io/>.

## [Unsupervised Visual Representation Learning via Mutual Information Regularized Assignment](#)

- Dong Hoon Lee · Sungik Choi · Hyunwoo Kim · Sae-Young Chung
- abstract@[open-review](#): This paper proposes Mutual Information Regularized Assignment (MIRA), a pseudo-labeling algorithm for unsupervised representation learning inspired by information maximization. We formulate online pseudo-labeling as an optimization problem to find pseudo-labels that

maximize the mutual information between the label and data while being close to a given model probability. We derive a fixed-point iteration method and prove its convergence to the optimal solution. In contrast to baselines, MIRA combined with pseudo-label prediction enables a simple yet effective clustering-based representation learning without incorporating extra training techniques or artificial constraints such as sampling strategy, equipartition constraints, etc. With relatively small training epochs, representation learned by MIRA achieves state-of-the-art performance on various downstream tasks, including the linear/ $\| \cdot \|_k$ -NN evaluation and transfer learning. Especially, with only 400 epochs, our method applied to ImageNet dataset with ResNet-50 architecture achieves 75.5% linear evaluation accuracy.

## [Differentiable hierarchical and surrogate gradient search for spiking neural networks](#)

- Kaiwei Che · Kaixuan Zhang · Liziwei Leng · Jianguo Zhang · Qinghu Meng · Jie Cheng · Qinghai Guo · Jianxing Liao
- abstract@[open-review](#): Spiking neural network (SNN) has been viewed as a potential candidate for the next generation of artificial intelligence, with appealing characteristics such as sparse computation and inherent temporal dynamics. By adopting sophisticated architectures of deep artificial neural networks (ANNs) and approximating backpropagation with surrogate gradient (SG) algorithms, deep SNNs are achieving competitive performances as their artificial counterparts in benchmark machine learning tasks such as image classification. However, successful architectures of ANNs are not necessary ideal for SNN and when tasks become more diverse effective architectural variations could be critical. To this end, we develop a differentiable hierarchical search framework for spiking neurons, where spike-based computation is realized on both the cell and the layer level search space. Based on this framework, we find effective SNN architectures under limited computation cost. During the training of SNN, a suboptimal SG function could lead to poor approximations of true gradients, making the network enter certain local minima. To address this problem, we propose a differentiable surrogate gradient search method where the SG function can be efficiently optimized locally in parallel. Our models achieve state-of-the-art performances on image classification. On event-based deep stereo, our method surpasses the accuracy of specially designed ANNs with 26  $\times$  lower energy cost ( $\$6.7\text{mJ}$ ), demonstrating the advantage of SNN in processing highly sparse and dynamic signals.

## [Could Giant Pre-trained Image Models Extract Universal Representations?](#)

- Yutong Lin · Ze Liu · Zheng Zhang · Han Hu · Nanning Zheng · Stephen Lin · Yue Cao
- abstract@[open-review](#): Frozen pretrained models have become a viable alternative to the pretraining-then-finetuning paradigm for transfer learning. However, with frozen models there are relatively few parameters available for adapting to downstream tasks, which is problematic in computer vision where tasks vary significantly in input/output format and the type of information that is of value. In this paper, we present a study of frozen pretrained models when applied to diverse and representative computer vision tasks, including object detection, semantic segmentation and video action recognition. From this empirical analysis, our work answers the questions of what pretraining task fits best with this frozen setting, how to make the frozen setting more flexible to various downstream tasks, and the effect of larger model sizes. We additionally examine the upper bound of performance using a giant frozen pretrained model with 3 billion parameters (SwinV2-G) and find that it reaches competitive performance on a varied set of major benchmarks with only one shared frozen base network: 60.0 box mAP and 52.2 mask mAP on COCO object detection test-dev, 57.6 val mIoU on ADE20K semantic segmentation, and 79.6 top-1 accuracy on Kinetics-400 action recognition. With this work, we hope to bring greater attention to this promising path of freezing pretrained image models.

## [On Learning and Refutation in Noninteractive Local Differential Privacy](#)

- Alexander Edmonds · Aleksandar Nikolov · Toniann Pitassi
- abstract@[open-review](#): We study two basic statistical tasks in non-interactive local differential privacy (LDP): *learning* and *refutation*: learning requires finding a concept that best fits an unknown target function (from labelled samples drawn from a distribution), whereas refutation requires distinguishing between data distributions that are well-correlated with some concept in the class, versus distributions where the labels are random. Our main result is a complete characterization of the sample complexity of agnostic PAC learning for non-interactive LDP protocols. We show that the optimal sample complexity for any concept class is captured by the approximate  $\|\gamma_2\|$  norm of a natural matrix associated with the class. Combined with previous work, this gives an *equivalence* between agnostic learning and refutation in the agnostic setting.

## [Margin-Based Few-Shot Class-Incremental Learning with Class-Level Overfitting Mitigation](#)

- Yixiong Zou · Shanghang Zhang · Yuhua Li · Ruixuan Li
- abstract@[open-review](#): Few-shot class-incremental learning (FSCIL) is designed to incrementally recognize novel classes with only few training samples after the (pre-)training on base classes with sufficient samples, which focuses on both base-class performance and novel-class generalization. A well known modification to the base-class training is to apply a margin to the base-class classification. However, a dilemma exists that we can hardly achieve both good base-class performance and novel-class generalization simultaneously by applying the margin during the base-class training, which is still under explored. In this paper, we study the cause of such dilemma for FSCIL. We first interpret this dilemma as a class-level overfitting (CO) problem from the aspect of pattern learning, and then find its cause lies in the easily-satisfied constraint of learning margin-based patterns. Based on the analysis, we propose a novel margin-based FSCIL method to mitigate the CO problem by providing the pattern learning process with extra constraint from the margin-based patterns themselves. Extensive experiments on CIFAR100, Caltech-USCD Birds-200-2011 (CUB200), and miniImageNet demonstrate that the proposed method effectively mitigates the CO problem and achieves state-of-the-art performance.

## [Toward a realistic model of speech processing in the brain with self-supervised learning](#)

- Juliette MILLET · Charlotte Caucheteux · pierre orhan · Yves Boubenec · Alexandre Gramfort · Ewan Dunbar · Christophe Pallier · Jean-Remi King
- abstract@[open-review](#): Several deep neural networks have recently been shown to generate activations similar to those of the brain in response to the same input. These algorithms, however, remain largely implausible: they require (1) extraordinarily large amounts of data, (2) unobtainable supervised labels, (3) textual rather than raw sensory input, and / or (4) implausibly large memory (e.g. thousands of contextual words). These elements highlight the need to identify algorithms that, under these limitations, would suffice to account for both behavioral and brain responses. Focusing on the issue of speech processing, we here hypothesize that self-supervised algorithms trained on the raw waveform constitute a promising candidate. Specifically, we compare a recent self-supervised architecture, Wav2Vec 2.0, to the brain activity of 412 English, French, and Mandarin individuals recorded with functional Magnetic Resonance Imaging (fMRI), while they listened to  $\approx 1$  h of audio books. Our results are four-fold. First, we show that this algorithm learns brain-like representations with as little as 600 hours of unlabelled speech -- a quantity comparable to what infants can be exposed to during language acquisition. Second, its functional hierarchy aligns with the cortical hierarchy of speech processing. Third, different training regimes reveal a functional specialization akin to the cortex: Wav2Vec 2.0 learns sound-generic, speech-specific and language-specific representations similar to those of the prefrontal and temporal cortices. Fourth, we confirm the similarity of this specialization with the behavior of 386 additional participants. These elements, resulting from the largest neuroimaging benchmark to date, show how self-supervised learning can account for a rich organization of speech processing in the brain, and thus delineate a path to identify the laws of language acquisition which shape the human brain.

## [Learning Tractable Probabilistic Models from Inconsistent Local Estimates](#)

- Shasha Jin · Vasundhara Komaragiri · Tahrima Rahman · Vibhav Gogate
- abstract@[open-review](#): Tractable probabilistic models or probabilistic circuits which admit exact linear time computation of either posterior marginal probabilities or most probable explanations (or both) are often preferred in practice over intractable models such as Bayesian and Markov networks. This

is because although tractable models are slightly inferior to the intractable models in terms of goodness-of-fit measures, they do not use approximate inference at prediction time and as a result exhibit superior predictive performance. In this paper, we consider the problem of improving a tractable model using local probability estimates for a small subset of variables (given observations) that are either available from experts or via an external process. The key idea in our approach is to update the parameters of the existing model via a gradient descent procedure that seeks to minimize a convex combination of two quantities: one that enforces closeness via KL divergence to the local estimates and another that enforces closeness to the given model. We show that although the gradients are NP-hard to compute on arbitrary Bayesian and Markov networks, they can be efficiently computed over tractable models. We show via experiments that our approach yields tractable models that are significantly superior to the ones learned from data alone even when the local estimates have large error.

## [Non-Linear Coordination Graphs](#)

- Yipeng Kang · Tonghan Wang · Qianlan Yang · Chongjie Zhang
- abstract@[open-review](#): Value decomposition multi-agent reinforcement learning methods learn the global value function as a mixing of each agent's individual utility functions. Coordination graphs (CGs) represent a higher-order decomposition by incorporating pairwise payoff functions and thus is supposed to have a more powerful representational capacity. However, CGs decompose the global value function linearly over local value functions, severely limiting the complexity of the value function class that can be represented. In this paper, we propose the first non-linear coordination graph by extending CG value decomposition beyond the linear case. One major challenge is to conduct greedy action selections in this new function class to which commonly adopted DCOP algorithms are no longer applicable. We study how to solve this problem when mixing networks with LeakyReLU activation are used. An enumeration method with a global optimality guarantee is proposed and motivates an efficient iterative optimization method with a local optimality guarantee. We find that our method can achieve superior performance on challenging multi-agent coordination tasks like MACO.

## [Homomorphic Matrix Completion](#)

- Zechu Li · Xiao-Yang Liu · Xiaodong Wang
- abstract@[open-review](#): In recommendation systems, global positioning, system identification and mobile social networks, it is a fundamental routine that a server completes a low-rank matrix from an observed subset of its entries. However, sending data to a cloud server raises up the data privacy concern due to eavesdropping attacks and the single-point failure problem, e.g., the Netflix prize contest was canceled after a privacy lawsuit. In this paper, we propose a homomorphic matrix completion algorithm for privacy-preserving data completion. First, we formulate a \textit{homomorphic matrix completion} problem where a server performs matrix completion on ciphertexts, and propose an encryption scheme that is fast and easy to implement. Secondly, we prove that the proposed scheme satisfies the \textit{homomorphism property} that decrypting the recovered matrix on ciphertexts will obtain the target complete matrix in plaintext. Thirdly, we prove that the proposed scheme satisfies an  $\$O(\sqrt{10}\{n_1^3 n_2\})\$$ -differential privacy property. While with similar level of privacy guarantee, we reduce the best-known error bound  $\$O(\sqrt{10}\{n_1^3 n_2\})\$$  to EXACT recovery at a price of more samples. Finally, on numerical data and real-world data, we show that both homomorphic nuclear-norm minimization and alternating minimization algorithms achieve accurate recoveries on ciphertexts, verifying the homomorphism property.

## [Effective Backdoor Defense by Exploiting Sensitivity of Poisoned Samples](#)

- Weixin Chen · Baoyuan Wu · Haoqian Wang
- abstract@[open-review](#): Poisoning-based backdoor attacks are serious threat for training deep models on data from untrustworthy sources. Given a backdoored model, we observe that the feature representations of poisoned samples with trigger are more sensitive to transformations than those of clean samples. It inspires us to design a simple sensitivity metric, called \textit{feature consistency towards transformations (FCT)}, to distinguish poisoned samples from clean samples in the untrustworthy training set. Moreover, we propose two effective backdoor defense methods. Built upon a sample-discrimination module utilizing the FCT metric, the first method trains a secure model from scratch using a two-stage secure training module. And the second method removes backdoor from a backdoored model with a backdoor removal module which alternatively unlearns the distinguished poisoned samples and relearns the distinguished clean samples. Extensive results on three benchmark datasets demonstrate the superior defense performance against eight types of backdoor attacks, to state-of-the-art backdoor defenses.

## [LION: Latent Point Diffusion Models for 3D Shape Generation](#)

- xiaohui zeng · Arash Vahdat · Francis Williams · Zan Gojcic · Or Litany · Sanja Fidler · Karsten Kreis
- abstract@[open-review](#): Denoising diffusion models (DDMs) have shown promising results in 3D point cloud synthesis. To advance 3D DDMs and make them useful for digital artists, we require (i) high generation quality, (ii) flexibility for manipulation and applications such as conditional synthesis and shape interpolation, and (iii) the ability to output smooth surfaces or meshes. To this end, we introduce the hierarchical Latent Point Diffusion Model (LION) for 3D shape generation. LION is set up as a variational autoencoder (VAE) with a hierarchical latent space that combines a global shape latent representation with a point-structured latent space. For generation, we train two hierarchical DDMs in these latent spaces. The hierarchical VAE approach boosts performance compared to DDMs that operate on point clouds directly, while the point-structured latents are still ideally suited for DDM-based modeling. Experimentally, LION achieves state-of-the-art generation performance on multiple ShapeNet benchmarks. Furthermore, our VAE framework allows us to easily use LION for different relevant tasks without re-training the latent DDMs: We show that LION excels at multimodal shape denoising and voxel-conditioned synthesis. We also demonstrate shape autoencoding and latent shape interpolation, and we augment LION with modern surface reconstruction techniques to generate smooth 3D meshes. We hope that LION provides a powerful tool for artists working with 3D shapes due to its high-quality generation, flexibility, and surface reconstruction.

## [Evolving Zero Cost Proxies For Neural Architecture Scoring](#)

- Yash Akhauri · Juan Munoz · Nilesh Jain · Ravishankar Iyer
- abstract@[open-review](#): Neural Architecture Search (NAS) has significantly improved productivity in the design and deployment of neural networks (NN). As NAS typically evaluates multiple models by training them partially or completely, the improved productivity comes at the cost of significant carbon footprint. To alleviate this expensive training routine, zero-shot/cost proxies analyze an NN at initialization to generate a score, which correlates highly with its true accuracy. Zero-cost proxies are currently designed by experts conducting multiple cycles of empirical testing on possible algorithms, data-sets, and neural architecture design spaces. This lowers productivity and is an unsustainable approach towards zero-cost proxy design as deep learning use-cases diversify in nature. Additionally, existing zero-cost proxies fail to generalize across neural architecture design spaces. In this paper, we propose a genetic programming framework to automate the discovery of zero-cost proxies for neural architecture scoring. Our methodology efficiently discovers an interpretable and generalizable zero-cost proxy that gives state of the art score-accuracy correlation on all data-sets and search spaces of NASBench-201 and Network Design Spaces (NDS). We believe that this research indicates a promising direction towards automatically discovering zero-cost proxies that can work across network architecture design spaces, data-sets, and tasks.

## [Learning Chaotic Dynamics in Dissipative Systems](#)

- Zongyi Li · Miguel Liu-Schiaffini · Nikola Kovachki · Kamyar Azizzadenesheli · Burigede Liu · Kaushik Bhattacharya · Andrew Stuart · Anima Anandkumar
- abstract@[open-review](#): Chaotic systems are notoriously challenging to predict because of their sensitivity to perturbations and errors due to time stepping. Despite this unpredictable behavior, for many dissipative systems the statistics of the long term trajectories are governed by an invariant measure

supported on a set, known as the global attractor; for many problems this set is finite dimensional, even if the state space is infinite dimensional. For Markovian systems, the statistical properties of long-term trajectories are uniquely determined by the solution operator that maps the evolution of the system over arbitrary positive time increments. In this work, we propose a machine learning framework to learn the underlying solution operator for dissipative chaotic systems, showing that the resulting learned operator accurately captures short-time trajectories and long-time statistical behavior. Using this framework, we are able to predict various statistics of the invariant measure for the turbulent Kolmogorov Flow dynamics with Reynolds numbers up to \$5000\$.

## [Visual Clues: Bridging Vision and Language Foundations for Image Paragraph Captioning](#)

- Yujia Xie · Luwei Zhou · Xiyang Dai · Lu Yuan · Nguyen Bach · Ce Liu · Michael Zeng
- abstract@[open-review](#): People say, "A picture is worth a thousand words". Then how can we get the rich information out of the image? We argue that by using visual clues to bridge large pretrained vision foundation models and language models, we can do so without any extra cross-modal training. Thanks to the strong zero-shot capability of foundation models, we start by constructing a rich semantic representation of the image (e.g., image tags, object attributes / locations, captions) as a structured textual prompt, called visual clues, using a vision foundation model. Based on visual clues, we use large language model to produce a series of comprehensive descriptions for the visual content, which is then verified by the vision model again to select the candidate that aligns best with the image. We evaluate the quality of generated descriptions by quantitative and qualitative measurement. The results demonstrate the effectiveness of such a structured semantic representation.

## [Agreement-on-the-line: Predicting the Performance of Neural Networks under Distribution Shift](#)

- Christina Baek · Yiding Jiang · Aditi Raghunathan · J. Zico Kolter
- abstract@[open-review](#): Recently, Miller et al. showed that a model's in-distribution (ID) accuracy has a strong linear correlation with its out-of-distribution (OOD) accuracy, on several OOD benchmarks, a phenomenon they dubbed ``accuracy-on-the-line''. While a useful tool for model selection (i.e., the model most likely to perform the best OOD is the one with highest ID accuracy), this fact does not help to estimate the actual OOD performance of models without access to a labeled OOD validation set. In this paper, we show a similar surprising phenomena also holds for the agreement between pairs of neural network classifiers: whenever accuracy-on-the-line holds, we observe that the OOD agreement between the predictions of any two pairs of neural networks (with potentially different architectures) also observes a strong linear correlation with their ID agreement. Furthermore, we observe that the slope and bias of OOD vs ID agreement closely matches that of OOD vs ID accuracy. This phenomenon which we call agreement-on-the-line, has important practical applications: without any labeled data, we can predict the OOD accuracy of classifiers, since OOD agreement can be estimated with just unlabeled data. Our prediction algorithm outperforms previous methods both in shifts where agreement-on-the-line holds and, surprisingly, when accuracy is not on the line. This phenomenon also provides new insights into neural networks: unlike accuracy-on-the-line, agreement-on-the-line only appears to hold for neural network classifiers.

## [Functional Indirection Neural Estimator for Better Out-of-distribution Generalization](#)

- Kha Pham · Thai Hung Le · Man Ngo · Truyen Tran
- abstract@[open-review](#): The capacity to achieve out-of-distribution (OOD) generalization is a hallmark of human intelligence and yet remains out of reach for machines. This remarkable capability has been attributed to our abilities to make conceptual abstraction and analogy, and to a mechanism known as indirection, which binds two representations and uses one representation to refer to the other. Inspired by these mechanisms, we hypothesize that OOD generalization may be achieved by performing analogy-making and indirection in the functional space instead of the data space as in current methods. To realize this, we design FINE (Functional Indirection Neural Estimator), a neural framework that learns to compose functions that map data input to output on-the-fly. FINE consists of a backbone network and a trainable semantic memory of basis weight matrices. Upon seeing a new input-output data pair, FINE dynamically constructs the backbone weights by mixing the basis weights. The mixing coefficients are indirectly computed through querying a separate corresponding semantic memory using the data pair. We demonstrate empirically that FINE can strongly improve out-of-distribution generalization on IQ tasks that involve geometric transformations. In particular, we train FINE and competing models on IQ tasks using images from the MNIST, Omniglot and CIFAR100 datasets and test on tasks with unseen image classes from one or different datasets and unseen transformation rules. FINE not only achieves the best performance on all tasks but also is able to adapt to small-scale data scenarios.

## [Neural-Symbolic Entangled Framework for Complex Query Answering](#)

- Zezhong Xu · Wen Zhang · Peng Ye · Hui Chen · Huajun Chen
- abstract@[open-review](#): Answering complex queries over knowledge graphs (KG) is an important yet challenging task because of the KG incompleteness issue and cascading errors during reasoning. Recent query embedding (QE) approaches embed the entities and relations in a KG and the first-order logic (FOL) queries into a low dimensional space, making the query can be answered by dense similarity searching. However, previous works mainly concentrate on the target answers, ignoring intermediate entities' usefulness, which is essential for relieving the cascading error problem in logical query answering. In addition, these methods are usually designed with their own geometric or distributional embeddings to handle logical operators like union, intersection, and negation, with the sacrifice of the accuracy of the basic operator -- projection, and they could not absorb other embedding methods to their models. In this work, we propose a Neural and Symbolic Entangled framework (ENeSy) for complex query answering, which enables the neural and symbolic reasoning to enhance each other to alleviate the cascading error and KG incompleteness. The projection operator in ENeSy could be any embedding method with the capability of link prediction, and the other FOL operators are handled without parameters. With both neural and symbolic reasoning results contained, ENeSy answers queries in ensembles. We evaluate ENeSy on complex query answering benchmarks, and ENeSy achieves the state-of-the-art, especially in the setting of training model only with the link prediction task.

## [Memorization Without Overfitting: Analyzing the Training Dynamics of Large Language Models](#)

- Kushal Tirumala · Aram Markosyan · Luke Zettlemoyer · Armen Aghajanyan
- abstract@[open-review](#): Despite their wide adoption, the underlying training and memorization dynamics of very large language models is not well understood. We empirically study exact memorization in causal and masked language modeling, across model sizes and throughout the training process. We measure the effects of dataset size, learning rate, and model size on memorization, finding that larger language models memorize training data faster across all settings. Surprisingly, we show that larger models can memorize a larger portion of the data before over-fitting and tend to forget less throughout the training process. We also analyze the memorization dynamics of different parts of speech and find that models memorize nouns and numbers first; we hypothesize and provide empirical evidence that nouns and numbers act as a unique identifier for memorizing individual training examples. Together, these findings present another piece of the broader puzzle of trying to understand what actually improves as models get bigger.

## [Learning-based Manipulation Planning in Dynamic Environments Using GNNs and Temporal Encoding](#)

- Ruipeng Zhang · Chennng Yu · Jingkai Chen · Chuchu Fan · Sicun Gao
- abstract@[open-review](#): Learning-based approaches have shown promising performance for improving the efficiency of motion planning in robot manipulation problems, but mostly in the setting of static environments. For the more challenging problem of motion planning in dynamic environments, such as for multi-arm assembly tasks or human-robot interaction, motion planners need to consider the trajectories of the dynamic obstacles, and reason about the temporal-spatial interactions between the ego-arm and the other objects. We propose a GNN-based neural architecture that involves temporal encoding, and use imitation learning with data aggregation procedures for learning both the embedding and edge prioritization policies. Experiments show

that the learning-based approach can significantly accelerate online planning in comparison to state-of-the-art complete dynamic planning algorithms. The proposed methods can reduce costly collision checking operations by more than 1000x, thus reducing the online planning time by over 95%, while also achieving high success rate on hard instances.

## [Zonotope Domains for Lagrangian Neural Network Verification](#)

- Matt Jordan · Jonathan Hayase · Alex Dimakis · Sewoong Oh
- abstract@[open-review](#): Neural network verification aims to provide provable bounds for the output of a neural network for a given input range. Notable prior works in this domain have either generated bounds using abstract domains, which preserve some dependency between intermediate neurons in the network; or framed verification as an optimization problem and solved a relaxation using Lagrangian methods. A key drawback of the latter technique is that each neuron is treated independently, thereby ignoring important neuron interactions. We provide an approach that merges these two threads and uses zonotopes within a Lagrangian decomposition. Crucially, we can decompose the problem of verifying a deep neural network into the verification of many 2-layer neural networks. While each of these problems is provably hard, we provide efficient relaxation methods that are amenable to efficient dual ascent procedures. Our technique yields bounds that improve upon both linear programming and Lagrangian-based verification techniques in both time and bound tightness.

## [Approaching Quartic Convergence Rates for Quasi-Stochastic Approximation with Application to Gradient-Free Optimization](#)

- Caio Kalil Lauand · Sean Meyn
- abstract@[open-review](#): Stochastic approximation is a foundation for many algorithms found in machine learning and optimization. It is in general slow to converge: the mean square error vanishes as  $\mathcal{O}(n^{-1})$ . A deterministic counterpart known as quasi-stochastic approximation (QSA) is a viable alternative in many applications, including gradient free optimization and reinforcement learning. It was assumed in recent prior research that the optimal achievable convergence rate is  $\mathcal{O}(n^{-2})$ . It is shown in this paper that through design it is possible to obtain far faster convergence, of order  $\mathcal{O}(n^{-4+\delta})$ , with  $\delta > 0$  arbitrary. Two acceleration techniques are introduced for the first time to achieve this rate of convergence. The theory is also specialized within the context of gradient free optimization, and tested on standard benchmarks. The main results are based on a combination of recent results from number theory and techniques adapted from stochastic approximation theory.

## [Non-convex online learning via algorithmic equivalence](#)

- Udaya Ghai · Zhou Lu · Elad Hazan
- abstract@[open-review](#): We study an algorithmic equivalence technique between nonconvex gradient descent and convex mirror descent. We start by looking at a harder problem of regret minimization in online non-convex optimization. We show that under certain geometric and smoothness conditions, online gradient descent applied to non-convex functions is an approximation of online mirror descent applied to convex functions under reparameterization. In continuous time, the gradient flow with this reparameterization was shown to be \emph{exactly} equivalent to continuous-time mirror descent by Amid and Warmuth, but theory for the analogous discrete time algorithms is left as an open problem. We prove an  $\mathcal{O}(T^{\frac{1}{3}})$  regret bound for non-convex online gradient descent in this setting, answering this open problem. Our analysis is based on a new and simple algorithmic equivalence method. \end{abstract}

## [Solving Quantitative Reasoning Problems with Language Models](#)

- Aitor Lewkowycz · Anders Andreassen · Vinay Ramasesh · Henryk Michalewski · David Dohan · Cem Anil · Ambrose Sloane · Imanol Schlag · Theo Gutman-Solo · Yuhuai Wu · Ethan Dyer · Guy Gur-Ari · Behnam Neyshabur · Vedant Misra
- abstract@[open-review](#): Language models have achieved remarkable performance on a wide range of tasks that require natural language understanding. Nevertheless, state-of-the-art models have generally struggled with tasks that require quantitative reasoning, such as solving mathematics, science, and engineering questions at the college level. To help close this gap, we introduce Minerva, a large language model pretrained on general natural language data and further trained on technical content. The model achieves strong performance in a variety of evaluations, including state-of-the-art performance on the MATH dataset. We also evaluate our model on over two hundred undergraduate-level problems in physics, biology, chemistry, economics, and other sciences that require quantitative reasoning, and find that the model can correctly answer nearly a quarter of them.

## [Detection and Localization of Changes in Conditional Distributions](#)

- Lizhen Nie · Dan Nicolae
- abstract@[open-review](#): We study the change point problem that considers alterations in the conditional distribution of an inferential target on a set of covariates. This paired data scenario is in contrast to the standard setting where a sequentially observed variable is analyzed for potential changes in the marginal distribution. We propose new methodology for solving this problem, by starting from a simpler task that analyzes changes in conditional expectation, and generalizing the tools developed for that task to conditional distributions. Large sample properties of the proposed statistics are derived. In empirical studies, we illustrate the performance of the proposed method against baselines adapted from existing tools. Two real data applications are presented to demonstrate its potential.

## [Exploring through Random Curiosity with General Value Functions](#)

- Aditya Ramesh · Louis Kirsch · Sjoerd van Steenkiste · Jürgen Schmidhuber
- abstract@[open-review](#): Efficient exploration in reinforcement learning is a challenging problem commonly addressed through intrinsic rewards. Recent prominent approaches are based on state novelty or variants of artificial curiosity. However, directly applying them to partially observable environments can be ineffective and lead to premature dissipation of intrinsic rewards. Here we propose random curiosity with general value functions (RC-GVF), a novel intrinsic reward function that draws upon connections between these distinct approaches. Instead of only using only the current observation's novelty or a curiosity bonus for failing to predict precise environment dynamics, RC-GVF derives intrinsic rewards through the task of predicting temporally extended general value functions. We demonstrate that this improves exploration in a hard-exploration diabolical lock problem. Further, RC-GVF significantly outperforms previous methods in the absence of ground-truth episodic counts in the partially observable MiniGrid environments. Panoramic observations on MiniGrid further boost RC-GVF's performance such that it is competitive to baselines exploiting episodic counts.

## [SIREN: Shaping Representations for OOD Detection](#)

- Xuefeng Du · Gabriel Gozum · Yifei Ming · Yixuan Li
- abstract@[open-review](#): Out-of-distribution (OOD) detection is indispensable for deploying machine learning models in the wild. Distance-based OOD detection methods are promising, but often suffer from discrepancies between the distributions learned in training vs. the distributional assumptions made in testing. This paper bridges the gap by addressing two key challenges---representation learning and OOD detection---in one coherent framework. Our proposed framework SIREN contributes two novel components: (1) a trainable loss function that shapes the representations into a mixture of von Mises-Fisher (vMF) distributions on the unit hypersphere, and (2) a test-time OOD detection score leveraging the learned vMF distributions. Unlike previous works, the two components in our framework enjoy strong mathematical compatibility with each other, under a unified distributional model. SIREN

achieves competitive performance on both the recent detection transformers and CNN-based models, improving the AUROC by over 10% compared to the previous best method on detection transformers.

## [Shape And Structure Preserving Differential Privacy](#)

- Carlos Soto · Karthik Bharath · Matthew Reimherr · Aleksandra Slavković‡
- abstract@[open-review](#): It is common for data structures such as images and shapes of 2D objects to be represented as points on a manifold. The utility of a mechanism to produce sanitized differentially private estimates from such data is intimately linked to how compatible it is with the underlying structure and geometry of the space. In particular, as recently shown, utility of the Laplace mechanism on a positively curved manifold, such as Kendall's 2D shape space, is significantly influenced by the curvature. Focusing on the problem of sanitizing the Fréchet mean of a sample of points on a manifold, we exploit the characterization of the mean as the minimizer of an objective function comprised of the sum of squared distances and develop a K-norm gradient mechanism on Riemannian manifolds that favors values that produce gradients close to the zero of the objective function. For the case of positively curved manifolds, we describe how using the gradient of the squared distance function offers better control over sensitivity than the Laplace mechanism, and demonstrate this numerically on a dataset of shapes of corpus callosum. Further illustrations of the mechanism's utility on a sphere and the manifold of symmetric positive definite matrices are also presented.

## [Bellman Residual Orthogonalization for Offline Reinforcement Learning](#)

- Andrea Zanette · Martin J Wainwright
- abstract@[open-review](#): We study a reinforcement learning principle that approximates the Bellman equations by enforcing their validity only along an user-defined space of test functions. Focusing on applications to model-free offline RL with function approximation, we exploit this principle to derive confidence intervals for off-policy evaluation, as well as to optimize over policies within a prescribed policy class. We prove an oracle inequality on our policy optimization procedure in terms of a trade-off between the value and uncertainty of an arbitrary comparator policy. Different choices of test function spaces allow us to tackle different problems within a common framework. We characterize the loss of efficiency in moving from on-policy to off-policy data using our procedures, and establish connections to concentrability coefficients studied in past work. We examine in depth the implementation of our methods with linear function approximation, and provide theoretical guarantees with polynomial-time implementations even when Bellman closure does not hold.

## [Reinforcement Learning with Non-Exponential Discounting](#)

- Matthias Schultheis · Constantin Rothkopf · Heinz Koeppl
- abstract@[open-review](#): Commonly in reinforcement learning (RL), rewards are discounted over time using an exponential function to model time preference, thereby bounding the expected long-term reward. In contrast, in economics and psychology, it has been shown that humans often adopt a hyperbolic discounting scheme, which is optimal when a specific task termination time distribution is assumed. In this work, we propose a theory for continuous-time reinforcement learning generalized to arbitrary discount functions. This formulation covers the case in which there is a random termination time. We derive a Hamilton-Jacobi-Bellman (HJB) equation characterizing the optimal policy and describe how it can be solved using a collocation method, which uses deep learning for function approximation. Further, we show how the inverse RL problem can be approached, in which one tries to recover properties of the discount function given decision data. We validate the applicability of our proposed approach on two simulated problems. Our approach opens the way for the analysis of human discounting in sequential decision-making tasks.

## [HUMANISE: Language-conditioned Human Motion Generation in 3D Scenes](#)

- Zan Wang · Yixin Chen · Tengyu Liu · Yixin Zhu · Wei Liang · Siyuan Huang
- abstract@[open-review](#): Learning to generate diverse scene-aware and goal-oriented human motions in 3D scenes remains challenging due to the mediocre characters of the existing datasets on Human-Scene Interaction (HSI); they only have limited scale/quality and lack semantics. To fill in the gap, we propose a large-scale and semantic-rich synthetic HSI dataset, denoted as HUMANISE, by aligning the captured human motion sequences with various 3D indoor scenes. We automatically annotate the aligned motions with language descriptions that depict the action and the individual interacting objects; e.g., sit on the armchair near the desk. HUMANISE thus enables a new generation task, language-conditioned human motion generation in 3D scenes. The proposed task is challenging as it requires joint modeling of the 3D scene, human motion, and natural language. To tackle this task, we present a novel scene-and-language conditioned generative model that can produce 3D human motions of the desirable action interacting with the specified objects. Our experiments demonstrate that our model generates diverse and semantically consistent human motions in 3D scenes.

## [A Simple Decentralized Cross-Entropy Method](#)

- Zichen Zhang · Jun Jin · Martin Jagersand · Jun Luo · Dale Schuurmans
- abstract@[open-review](#): Cross-Entropy Method (CEM) is commonly used for planning in model-based reinforcement learning (MBRL) where a centralized approach is typically utilized to update the sampling distribution based on only the top-\$k\$ operations' results on samples. In this paper, we show that such a centralized approach makes CEM vulnerable to local optima, thus impairing its sample efficiency. To tackle this issue, we propose Decentralized CEM (DecentCEM), a simple but effective improvement over classical CEM, by using an ensemble of CEM instances running independently from one another, and each performing a local improvement of its own sampling distribution. We provide both theoretical and empirical analysis to demonstrate the effectiveness of this simple decentralized approach. We empirically show that, compared to the classical centralized approach using either a single or even a mixture of Gaussian distributions, our DecentCEM finds the global optimum much more consistently thus improves the sample efficiency. Furthermore, we plug in our DecentCEM in the planning problem of MBRL, and evaluate our approach in several continuous control environments, with comparison to the state-of-art CEM based MBRL approaches (PETS and POPLIN). Results show sample efficiency improvement by simply replacing the classical CEM module with our DecentCEM module, while only sacrificing a reasonable amount of computational cost, which can be alleviated by using the advantage of parallelism of our ensemble method. Lastly, we conduct ablation studies for more in-depth analysis.

## [Evolution of Neural Tangent Kernels under Benign and Adversarial Training](#)

- Noel Loo · Ramin Hasani · Alexander Amini · Daniela Rus
- abstract@[open-review](#): Two key challenges facing modern deep learning is mitigating deep networks vulnerability to adversarial attacks, and understanding deep learning's generalization capabilities. Towards the first issue, many defense strategies have been developed, with the most common being Adversarial Training (AT). Towards the second challenge, one of the dominant theories that has emerged is the Neural Tangent Kernel (NTK) -- a characterization of neural network behavior in the infinite-width limit. In this limit, the kernel is frozen and the underlying feature map is fixed. In finite-widths however, there is evidence that feature learning happens at the earlier stages of the training (kernel learning) before a second phase where the kernel remains fixed (lazy training). While prior work has aimed at studying adversarial vulnerability through the lens of the frozen infinite-width NTK, there is no work which studies adversarial robustness of NTK during training. In this work, we perform an empirical study of the evolution of the NTK under standard and adversarial training, aiming to disambiguate the effect of adversarial training on kernel learning and lazy training. We find under adversarial training, the NTK rapidly converges to a different kernel (and feature map) than standard training. This new kernel provides adversarial robustness, even when non-robust training is performed on top of it. Furthermore, we find that adversarial training on top of a fixed kernel can yield a classifier with \$76.1\%\$ robust accuracy under PGD attacks with \$\epsilon = 4/255\$ on CIFAR-10.

## [Efficient Dataset Distillation using Random Feature Approximation](#)

- Noel Loo · Ramin Hasani · Alexander Amini · Daniela Rus
- abstract@[open-review](#): Dataset distillation compresses large datasets into smaller synthetic coresets which retain performance with the aim of reducing the storage and computational burden of processing the entire dataset. Today's best performing algorithm, \textit{Kernel Inducing Points} (KIP), which makes use of the correspondence between infinite-width neural networks and kernel-ridge regression, is prohibitively slow due to the exact computation of the neural tangent kernel matrix, scaling  $\mathcal{O}(|S|^2)$ , with  $|S|$  being the coreset size. To improve this, we propose a novel algorithm that uses a random feature approximation (RFA) of the Neural Network Gaussian Process (NNGP) kernel which reduces the kernel matrix computation to  $\mathcal{O}(|S|)$ . Our algorithm provides at least a 100-fold speedup over KIP and can run on a single GPU. Our new method, termed an RFA Distillation (RFAD), performs competitively with KIP and other dataset condensation algorithms in accuracy over a range of large-scale datasets, both in kernel regression and finite-width network training. We demonstrate the effectiveness of our approach on tasks involving model interpretability and privacy preservation.

## [Handcrafted Backdoors in Deep Neural Networks](#)

- Sanghyun Hong · Nicholas Carlini · Alexey Kurakin
- abstract@[open-review](#): When machine learning training is outsourced to third parties, \$backdoor\$ \$attacks\$ become practical as the third party who trains the model may act maliciously to inject hidden behaviors into the otherwise accurate model. Until now, the mechanism to inject backdoors has been limited to \$poisoning\$. We argue that a supply-chain attacker has more attack techniques available by introducing a \$handcrafted\$ attack that directly manipulates a model's weights. This direct modification gives our attacker more degrees of freedom compared to poisoning, and we show it can be used to evade many backdoor detection or removal defenses effectively. Across four datasets and four network architectures our backdoor attacks maintain an attack success rate above 96%. Our results suggest that further research is needed for understanding the complete space of supply-chain backdoor attacks.

## [A Neural Pre-Conditioning Active Learning Algorithm to Reduce Label Complexity](#)

- Seo Taek Kong · Soomin Jeon · Dongbin Na · Jaewon Lee · Hong-Seok Lee · Kyu-Hwan Jung
- abstract@[open-review](#): Deep learning (DL) algorithms rely on massive amounts of labeled data, and semi-supervised learning (SSL) and active learning (AL) algorithms have been designed to reduce this label complexity by leveraging unlabeled data or carefully acquiring labels. In this work, we primarily focus on designing an AL algorithm but first argue for a change in how AL algorithms are evaluated. Although unlabeled data is readily available in pool-based AL, experimental evaluations had typically compared the performance improvements of supervised learning (SL) as labels are incrementally acquired. Instead we argue that the enhancements of SSL performance with added labels should be used to evaluate AL algorithms to measure their label efficiency. Focusing on this objective and after surveying tools that can be used to this end, we propose a neural pre-conditioning (NPC) algorithm, based on a neural tangent kernel (NTK) analysis, that evaluates unlabeled data based on how they would contribute upon inclusion to the training set. Our algorithm uses uncertainty information captured by the networks gradients as argued by the recently-proposed BADGE algorithm but differs in how diversity is enforced. Furthermore, we prove that NPC improves downstream training landscape in the NTK regime with respect to properties known to correlate with generalization. Comparisons with other AL algorithms show that a state-of-the-art SSL algorithm coupled with NPC can achieve high performance using very few labeled data.

## [Improved Imaging by Invex Regularizers with Global Optima Guarantees](#)

- Samuel Pinilla · Tingting Mu · Neil Bourne · Jeyan Thiyyagalingam
- abstract@[open-review](#): Image reconstruction enhanced by regularizers, e.g., to enforce sparsity, low rank or smoothness priors on images, has many successful applications in vision tasks such as computer photography, biomedical and spectral imaging. It has been well accepted that non-convex regularizers normally perform better than convex ones in terms of the reconstruction quality. But their convergence analysis is only established to a critical point, rather than the global optima. To mitigate the loss of guarantees for global optima, we propose to apply the concept of invexity and provide the first list of proved invex regularizers for improving image reconstruction. Moreover, we establish convergence guarantees to global optima for various advanced image reconstruction techniques after being improved by such invex regularization. To the best of our knowledge, this is the first practical work applying invex regularization to improve imaging with global optima guarantees. To demonstrate the effectiveness of invex regularization, numerical experiments are conducted for various imaging tasks using benchmark datasets.

## [Data Distributional Properties Drive Emergent In-Context Learning in Transformers](#)

- Stephanie Chan · Adam Santoro · Andrew Lampinen · Jane Wang · Aaditya Singh · Pierre Richemond · James McClelland · Felix Hill
- abstract@[open-review](#): Large transformer-based models are able to perform in-context few-shot learning, without being explicitly trained for it. This observation raises the question: what aspects of the training regime lead to this emergent behavior? Here, we show that this behavior is driven by the distributions of the training data itself. In-context learning emerges when the training data exhibits particular distributional properties such as burstiness (items appear in clusters rather than being uniformly distributed over time) and having a large number of rarely occurring classes. In-context learning also emerges more strongly when item meanings or interpretations are dynamic rather than fixed. These properties are exemplified by natural language, but are also inherent to naturalistic data in a wide range of other domains. They also depart significantly from the uniform, i.i.d. training distributions typically used for standard supervised learning. In our initial experiments, we found that in-context learning traded off against more conventional weight-based learning, and models were unable to achieve both simultaneously. However, our later experiments uncovered that the two modes of learning could co-exist in a single model when it was trained on data following a skewed Zipfian distribution -- another common property of naturalistic data, including language. In further experiments, we found that naturalistic data distributions were only able to elicit in-context learning in transformers, and not in recurrent models. Our findings indicate how the transformer architecture works together with particular properties of the training data to drive the intriguing emergent in-context learning behaviour of large language models, and indicate how future work might encourage both in-context and in-weights learning in domains beyond language.

## [Learning to Break the Loop: Analyzing and Mitigating Repetitions for Neural Text Generation](#)

- Jin Xu · Xiaojiang Liu · Jianhao Yan · Deng Cai · Huayang Li · Jian Li
- abstract@[open-review](#): While large-scale neural language models, such as GPT2 and BART, have achieved impressive results on various text generation tasks, they tend to get stuck in undesirable sentence-level loops with maximization-based decoding algorithms (\textit{e.g.}, greedy search). This phenomenon is counter-intuitive since there are few consecutive sentence-level repetitions in the human corpus (\textit{e.g.}, 0.02% in WikiText-103). To investigate the underlying reasons for generating consecutive sentence-level repetitions, we study the relationship between the probability of repetitive tokens and their previous repetitions in context. Through our quantitative experiments, we find that 1) Models have a preference to repeat the previous sentence; 2) The sentence-level repetitions have a \textit{self-reinforcement effect}: the more times a sentence is repeated in the context, the higher the probability of continuing to generate that sentence; 3) The sentences with higher initial probabilities usually have a stronger self-reinforcement effect. Motivated by our findings, we propose a simple and effective training method \textit{DITTO} (Pseudo\underline{D}o-\textit{Repet}\underline{I}tition \textit{Penaliza}\underline{T}\underline{O}n), where the model learns to penalize probabilities of sentence-level repetitions from synthetic repetitive data. Although our method is motivated by mitigating repetitions, our experiments show that DITTO not only mitigates the repetition issue without sacrificing perplexity, but also achieves better generation quality. Extensive experiments on open-ended text generation (WikiText-103) and text summarization (CNN/DailyMail) demonstrate the generality and effectiveness of our method.

## [Controlled Sparsity via Constrained Optimization or: How I Learned to Stop Tuning Penalties and Love Constraints](#)

- Jose Gallego-Posada · Juan Ramirez · Akram Erraqabi · Yoshua Bengio · Simon Lacoste-Julien
- abstract@[open-review](#): The performance of trained neural networks is robust to harsh levels of pruning. Coupled with the ever-growing size of deep learning models, this observation has motivated extensive research on learning sparse models. In this work, we focus on the task of controlling the level of sparsity when performing sparse learning. Existing methods based on sparsity-inducing penalties involve expensive trial-and-error tuning of the penalty factor, thus lacking direct control of the resulting model sparsity. In response, we adopt a constrained formulation: using the gate mechanism proposed by Louizos et al. (2018), we formulate a constrained optimization problem where sparsification is guided by the training objective and the desired sparsity target in an end-to-end fashion. Experiments on CIFAR-10/100, TinyImageNet, and ImageNet using WideResNet and ResNet{18, 50} models validate the effectiveness of our proposal and demonstrate that we can reliably achieve pre-determined sparsity targets without compromising on predictive performance.

## [Nonparametric Uncertainty Quantification for Single Deterministic Neural Network](#)

- Nikita Kotelevskii · Aleksandr Artemenkov · Kirill Fedyanin · Fedor Noskov · Alexander Fishkov · Artem Shelmanov · Artem Vazhentsev · Aleksandr Petushko · Maxim Panov
- abstract@[open-review](#): This paper proposes a fast and scalable method for uncertainty quantification of machine learning models' predictions. First, we show the principled way to measure the uncertainty of predictions for a classifier based on Nadaraya-Watson's nonparametric estimate of the conditional label distribution. Importantly, the approach allows to disentangle explicitly \textit{aleatoric} and \textit{epistemic} uncertainties. The resulting method works directly in the feature space. However, one can apply it to any neural network by considering an embedding of the data induced by the network. We demonstrate the strong performance of the method in uncertainty estimation tasks on text classification problems and a variety of real-world image datasets, such as MNIST, SVHN, CIFAR-100 and several versions of ImageNet.

## [Boosting the Transferability of Adversarial Attacks with Reverse Adversarial Perturbation](#)

- Zeyu Qin · Yanbo Fan · Yi Liu · Li Shen · Yong Zhang · Jue Wang · Baoyuan Wu
- abstract@[open-review](#): Deep neural networks (DNNs) have been shown to be vulnerable to adversarial examples, which can produce erroneous predictions by injecting imperceptible perturbations. In this work, we study the transferability of adversarial examples, which is significant due to its threat to real-world applications where model architecture or parameters are usually unknown. Many existing works reveal that the adversarial examples are likely to overfit the surrogate model that they are generated from, limiting its transfer attack performance against different target models. To mitigate the overfitting of the surrogate model, we propose a novel attack method, dubbed reverse adversarial perturbation (RAP). Specifically, instead of minimizing the loss of a single adversarial point, we advocate seeking adversarial example located at a region with unified low loss value, by injecting the worst-case perturbation (the reverse adversarial perturbation) for each step of the optimization procedure. The adversarial attack with RAP is formulated as a min-max bi-level optimization problem. By integrating RAP into the iterative process for attacks, our method can find more stable adversarial examples which are less sensitive to the changes of decision boundary, mitigating the overfitting of the surrogate model. Comprehensive experimental comparisons demonstrate that RAP can significantly boost adversarial transferability. Furthermore, RAP can be naturally combined with many existing black-box attack techniques, to further boost the transferability. When attacking a real-world image recognition system, Google Cloud Vision API, we obtain 22% performance improvement of targeted attacks over the compared method.

## [Knowledge Distillation from A Stronger Teacher](#)

- Tao Huang · Shan You · Fei Wang · Chen Qian · Chang Xu
- abstract@[open-review](#): Unlike existing knowledge distillation methods focus on the baseline settings, where the teacher models and training strategies are not that strong and competing as state-of-the-art approaches, this paper presents a method dubbed DIST to distill better from a stronger teacher. We empirically find that the discrepancy of predictions between the student and a stronger teacher may tend to be fairly severer. As a result, the exact match of predictions in KL divergence would disturb the training and make existing methods perform poorly. In this paper, we show that simply preserving the relations between the predictions of teacher and student would suffice, and propose a correlation-based loss to capture the intrinsic inter-class relations from the teacher explicitly. Besides, considering that different instances have different semantic similarities to each class, we also extend this relational match to the intra-class level. Our method is simple yet practical, and extensive experiments demonstrate that it adapts well to various architectures, model sizes and training strategies, and can achieve state-of-the-art performance consistently on image classification, object detection, and semantic segmentation tasks.

## [Improved surface reconstruction using high-frequency details](#)

- Yiqun Wang · Ivan Skorokhodov · Peter Wonka
- abstract@[open-review](#): Neural rendering can be used to reconstruct implicit representations of shapes without 3D supervision. However, current neural surface reconstruction methods have difficulty learning high-frequency details of shapes, so that the reconstructed shapes are often oversmoothed. We propose a novel method to improve the quality of surface reconstruction in neural rendering. We follow recent work to model surfaces as signed distance fields. First, we offer a derivation to analyze the relationship between the signed distance function, the volume density, the transparency function, and the weighting function used in the volume rendering equation. Second, we observe that attempting to jointly encode high-frequency and low frequency components in a single signed distance function leads to unstable optimization. We propose to decompose the signed distance function in a base function and a displacement function together with a coarse-to-fine strategy to gradually increase the high-frequency details. Finally, we propose to use an adaptive strategy that enables the optimization to focus on improving certain regions near the surface where the signed distance fields have artifacts. Our qualitative and quantitative results show that our method can reconstruct high-frequency surface details and obtain better surface reconstruction quality than the current state of the art.

## [BooNTK: Convexifying Federated Learning using Bootstrapped Neural Tangent Kernels](#)

- Yaodong Yu · Alexander Wei · Sai Praneeth Karimireddy · Yi Ma · Michael Jordan
- abstract@[open-review](#): State-of-the-art federated learning methods can perform far worse than their centralized counterparts when clients have dissimilar data distributions. For neural networks, even when centralized SGD easily finds a solution that is simultaneously performant for all clients, current federated optimization methods fail to converge to a comparable solution. We show that this performance disparity can largely be attributed to optimization challenges presented by non-convexity. Specifically, we find that the early layers of the network do learn useful features, but the final layers fail to make use of them. That is, federated optimization applied to this non-convex problem distorts the learning of the final layers. Leveraging this observation, we propose a simple, two-stage training procedure Bootstrapped eNTK (BooNTK) to sidestep this issue: first, learn features using off-the-shelf methods (e.g. FedAvg); then, optimize a \textit{convexified} problem obtained from the network's empirical neural tangent kernel approximation. Our technique yields accuracy improvements of up to \$+36\%\$ on FMNIST and \$+37\%\$ on CIFAR10 when clients have dissimilar data.

## [Language Conditioned Spatial Relation Reasoning for 3D Object Grounding](#)

- Shizhe Chen · Pierre-Louis Guhur · Makarand Tapaswi · Cordelia Schmid · Ivan Laptev

- abstract@[open-review](#): Localizing objects in 3D scenes based on natural language requires understanding and reasoning about spatial relations. In particular, it is often crucial to distinguish similar objects referred by the text, such as "the left most chair" and "a chair next to the window". In this work we propose a language-conditioned transformer model for grounding 3D objects and their spatial relations. To this end, we design a spatial self-attention layer that accounts for relative distances and orientations between objects in input 3D point clouds. Training such a layer with visual and language inputs enables to disambiguate spatial relations and to localize objects referred by the text. To facilitate the cross-modal learning of relations, we further propose a teacher-student approach where the teacher model is first trained using ground-truth object labels, and then helps to train a student model using point cloud inputs. We perform ablation studies showing advantages of our approach. We also demonstrate our model to significantly outperform the state of the art on the challenging Nr3D, Sr3D and ScanRefer 3D object grounding datasets. Our code and pretrained models will become publicly available.

## [Revisiting Sparse Convolutional Model for Visual Recognition](#)

- xili dai Â· Mingyang Li Â· Pengyuan Zhai Â· Shengbang Tong Â· Xingjian Gao Â· Shao-Lun Huang Â· Zhihui Zhu Â· Chong You Â· Yi Ma
- abstract@[open-review](#): Despite strong empirical performance for image classification, deep neural networks are often regarded as ``black boxes'' and they are difficult to interpret. On the other hand, sparse convolutional models, which assume that a signal can be expressed by a linear combination of a few elements from a convolutional dictionary, are powerful tools for analyzing natural images with good theoretical interpretability and biological plausibility. However, such principled models have not demonstrated competitive performance when compared with empirically designed deep networks. This paper revisits the sparse convolutional modeling for image classification and bridges the gap between good empirical performance (of deep learning) and good interpretability (of sparse convolutional models). Our method uses differentiable optimization layers that are defined from convolutional sparse coding as drop-in replacements of standard convolutional layers in conventional deep neural networks. We show that such models have equally strong empirical performance on CIFAR-10, CIFAR-100 and ImageNet datasets when compared to conventional neural networks. By leveraging stable recovery property of sparse modeling, we further show that such models can be much more robust to input corruptions as well as adversarial perturbations in testing through a simple proper trade-off between sparse regularization and data reconstruction terms.

## [Category-Level 6D Object Pose Estimation in the Wild: A Semi-Supervised Learning Approach and A New Dataset](#)

- Yang Fu Â· Xiaolong Wang
- abstract@[open-review](#): 6D object pose estimation is one of the fundamental problems in computer vision and robotics research. While a lot of recent efforts have been made on generalizing pose estimation to novel object instances within the same category, namely category-level 6D pose estimation, it is still restricted in constrained environments given the limited number of annotated data. In this paper, we collect Wild6D, a new unlabeled RGBD object video dataset with diverse instances and backgrounds. We utilize this data to generalize category-level 6D object pose estimation in the wild with semi-supervised learning. We propose a new model, called Rendering for Pose estimation network RePoNet), that is jointly trained using the free ground-truths with the synthetic data, and a silhouette matching objective function on the real-world data. Without using any 3D annotations on real data, our method outperforms state-of-the-art methods on the previous dataset and our Wild6D test set (with manual annotations for evaluation) by a large margin. Our code and dataset will be made publicly available.

## [An Investigation into Whitening Loss for Self-supervised Learning](#)

- Xi Weng Â· Lei Huang Â· Lei Zhao Â· Rao Anwer Â· Salman Khan Â· Fahad Shahbaz Khan
- abstract@[open-review](#): A desirable objective in self-supervised learning (SSL) is to avoid feature collapse. Whitening loss guarantees collapse avoidance by minimizing the distance between embeddings of positive pairs under the conditioning that the embeddings from different views are whitened. In this paper, we propose a framework with an informative indicator to analyze whitening loss, which provides a clue to demystify several interesting phenomena as well as a pivoting point connecting to other SSL methods. We reveal that batch whitening (BW) based method do not impose whitening constraints on the embedding, but they only require the embedding to be full-rank. This full-rank constraint is also sufficient to avoid dimensional collapse. Based on our analysis, we propose a channel whitening with random group partition (CW-RGP), which exploits the advantages of BW-based method in preventing collapse and avoids their disadvantages for large batch size. Experimental results on ImageNet classification and COCO object detection reveal that the proposed CW-RGP possesses a promising potential for learning good representations.

## [ProcTHOR: Large-Scale Embodied AI Using Procedural Generation](#)

- Matt Deitke Â· Eli VanderBilt Â· Alvaro Herrasti Â· Winson Han Â· Luca Weihs Â· Kiana Ehsani Â· Jordi Salvador Â· Eric Kolve Â· Aniruddha Kembhavi Â· Roozbeh Mottaghi
- abstract@[open-review](#): Massive datasets and high-capacity models have driven many recent advancements in computer vision and natural language understanding. This work presents a platform to enable similar success stories in Embodied AI. We propose ProcTHOR, a framework for procedural generation of Embodied AI environments. ProcTHOR enables us to sample arbitrarily large datasets of diverse, interactive, customizable, and performant virtual environments to train and evaluate embodied agents across navigation, interaction, and manipulation tasks. We demonstrate the power and potential of ProcTHOR via a sample of 10,000 generated houses and a simple neural model. Models trained using only RGB images on ProcTHOR, with no explicit mapping and no human task supervision produce state-of-the-art results across 6 embodied AI benchmarks for navigation, rearrangement, and arm manipulation, including the presently running Habitat 2022, AI2-THOR Rearrangement 2022, and RoboTHOR challenges. We also demonstrate strong 0-shot results on these benchmarks, via pre-training on ProcTHOR with no fine-tuning on the downstream benchmark, often beating previous state-of-the-art systems that access the downstream training data.

## [Symmetry Teleportation for Accelerated Optimization](#)

- Bo Zhao Â· Nima Dehmamy Â· Robin Walters Â· Rose Yu
- abstract@[open-review](#): Existing gradient-based optimization methods update the parameters locally, in a direction that minimizes the loss function. We study a different approach, symmetry teleportation, that allows the parameters to travel a large distance on the loss level set, in order to improve the convergence speed in subsequent steps. Teleportation exploits parameter space symmetries of the optimization problem and transforms parameters while keeping the loss invariant. We derive the loss-invariant group actions for test functions and multi-layer neural networks, and prove a necessary condition of when teleportation improves convergence rate. We also show that our algorithm is closely related to second order methods. Experimentally, we show that teleportation improves the convergence speed of gradient descent and AdaGrad for several optimization problems including test functions, multi-layer regressions, and MNIST classification.

## [Generalizing Consistent Multi-Class Classification with Rejection to be Compatible with Arbitrary Losses](#)

- Yuzhou Cao Â· Tianchi Cai Â· Lei Feng Â· Lihong Gu Â· Jinjie GU Â· Bo An Â· Gang Niu Â· Masashi Sugiyama
- abstract@[open-review](#): \text{Classification with rejection} (CwR) refrains from making a prediction to avoid critical misclassification when encountering test samples that are difficult to classify. Though previous methods for CwR have been provided with theoretical guarantees, they are only compatible with certain loss functions, making them not flexible enough when the loss needs to be changed with the dataset in practice. In this paper, we derive a novel formulation for CwR that can be equipped with arbitrary loss functions while maintaining the theoretical guarantees. First, we show that \$K\$-class CwR is equivalent to a \$(K!+1)\$-class classification problem on the original data distribution with an augmented class, and propose an empirical risk minimization formulation to solve this problem with an estimation error bound. Then, we find necessary and sufficient conditions for the learning \text{consistency} of the surrogates constructed on our proposed formulation equipped with any classification-calibrated multi-class losses,

where consistency means the surrogate risk minimization implies the target risk minimization for CwR. Finally, experiments on benchmark datasets validate the effectiveness of our proposed method.

## [Multi-view Subspace Clustering on Topological Manifold](#)

- Shudong Huang · Hongjie Wu · Yazhou Ren · Ivor Tsang · Zenglin Xu · Jiancheng Lv · Wentao Feng
- abstract@[open-review](#): Multi-view subspace clustering aims to exploit a common affinity representation by means of self-expression. Plenty of works have been presented to boost the clustering performance, yet seldom considering the topological structure in data, which is crucial for clustering data on manifold. Orthogonal to existing works, in this paper, we argue that it is beneficial to explore the implied data manifold by learning the topological relationship between data points. Our model seamlessly integrates multiple affinity graphs into a consensus one with the topological relevance considered. Meanwhile, we manipulate the consensus graph by a connectivity constraint such that the connected components precisely indicate different clusters. Hence our model is able to directly obtain the final clustering result without reliance on any label discretization strategy as previous methods do. Experimental results on several benchmark datasets illustrate the effectiveness of the proposed model, compared to the state-of-the-art competitors over the clustering performance.

## [ReFactorGNNs: Revisiting Factorisation-based Models from a Message-Passing Perspective](#)

- Yihong Chen · Pushkar Mishra · Luca Franceschi · Pasquale Minervini · Pontus Lars Erik Saito Stenetorp · Sebastian Riedel
- abstract@[open-review](#): Factorisation-based Models (FMs), such as DistMult, have enjoyed enduring success for Knowledge Graph Completion (KGC) tasks, often outperforming Graph Neural Networks (GNNs). However, unlike GNNs, FMs struggle to incorporate node features and to generalise to unseen nodes in inductive settings. Our work bridges the gap between FMs and GNNs by proposing REFACTOR GNNS. This new architecture draws upon both modelling paradigms, which previously were largely thought of as disjoint. Concretely, using a message-passing formalism, we show how FMs can be cast as GNNs by reformulating the gradient descent procedure as message-passing operations, which forms the basis of our REFACTOR GNNS. Across a multitude of well-established KGC benchmarks, our REFACTOR GNNS achieve comparable transductive performance to FMs, and state-of-the-art inductive performance while using an order of magnitude fewer parameters.

## [Squeezeformer: An Efficient Transformer for Automatic Speech Recognition](#)

- Amir Gholami · Kurt Keutzer · Sehoon Kim · Nicholas Lee · Michael Mahoney · Jitendra Malik · Karttikeya Mangalam · Albert Shaw
- abstract@[open-review](#): The recently proposed Conformer model has become the de facto backbone model for various downstream speech tasks based on its hybrid attention-convolution architecture that captures both local and global features. However, through a series of systematic studies, we find that the Conformer architecture's design choices are not optimal. After reexamining the design choices for both the macro and micro-architecture of Conformer, we propose Squeezeformer which consistently outperforms the state-of-the-art ASR models under the same training schemes. In particular, for the macro-architecture, Squeezeformer incorporates (i) the Temporal U-Net structure %downsamples and upsamples speech frames, which reduces the cost of the multi-head attention modules on long sequences, and (ii) a simpler block structure of feed-forward module followed up by multi-head attention or convolution modules instead of the Macaron structure proposed in Conformer. Furthermore, for the micro-architecture, Squeezeformer (i) simplifies the activations in the convolutional block, (ii) removes redundant Layer Normalization operations, and (iii) incorporates an efficient depth-wise downsampling layer to efficiently sub-sample the input signal. Squeezeformer achieves state-of-the-art results of 7.5%, 6.5%, and 6.0% word-error-rate (WER) on Librispeech test-other without external language models, which are 3.1%, 1.4%, and 0.6% better than Conformer-CTC with the same number of FLOPs. Our code is open-sourced and available online.

## [On the consistent estimation of optimal Receiver Operating Characteristic \(ROC\) curve](#)

- Renxiong Liu · Yunzhang Zhu
- abstract@[open-review](#): Under a standard binary classification setting with possible model misspecification, we study the problem of estimating general Receiver Operating Characteristic (ROC) curve, which is an arbitrary set of false positive rate (FPR) and true positive rate (TPR) pairs. We formally introduce the notion of \textit{optimal ROC curve} over a general model space. It is argued that any ROC curve estimation methods implemented over the given model space should target the optimal ROC curve over that space. Three popular ROC curve estimation methods are then analyzed at the population level (i.e., when there are infinite number of samples) under both correct and incorrect model specification. Based on our analysis, they are all consistent when the surrogate loss function satisfies certain conditions and the given model space includes all measurable classifiers. Interestingly, some of these conditions are similar to those that are required to ensure classification consistency. When the model space is incorrectly specified, however, we show that only one method leads to consistent estimation of the ROC curve over the chosen model space. We present some numerical results to demonstrate the effects of model misspecification on the performance of various methods in terms of their ROC curve estimates.

## [Learning Recourse on Instance Environment to Enhance Prediction Accuracy](#)

- Lokesh N · Guntakanti Sai Koushik · Abir De · Sunita Sarawagi
- abstract@[open-review](#): Machine Learning models are often susceptible to poor performance on instances sampled from bad environments. For example, an image classifier could provide low accuracy on images captured under low lighting conditions. In high stake ML applications, such as AI-driven medical diagnostics, a better option could be to provide recourse in the form of alternative environment settings in which to recapture the instance for more reliable diagnostics. In this paper, we propose a model called {\em RecourseNet} that learns to apply recourse on the space of environments so that the recoursed instances are amenable to better predictions by the classifier. Learning to output optimal recourse is challenging because we do not assume access to the underlying physical process that generates the recoursed instances. Also, the optimal setting could be instance-dependent --- for example the best camera angle for object recognition could be a function of the object's shape. We propose a novel three-level training method that (a) Learns a classifier that is optimized for high performance under recourse, (b) Learns a recourse predictor when the training data may contain only limited instances under good environment settings, and (c) Triggers recourse selectively only when recourse is likely to improve classifier confidence.

## [ULNeF: Untangled Layered Neural Fields for Mix-and-Match Virtual Try-On](#)

- Igor Santesteban · Miguel Otaduy · Nils Thuerey · Dan Casas
- abstract@[open-review](#): Recent advances in neural models have shown great results for virtual try-on (VTO) problems, where a 3D representation of a garment is deformed to fit a target body shape. However, current solutions are limited to a single garment layer, and cannot address the combinatorial complexity of mixing different garments. Motivated by this limitation, we investigate the use of neural fields for mix-and-match VTO, and identify and solve a fundamental challenge that existing neural-field methods cannot address: the interaction between layered neural fields. To this end, we propose a neural model that untangles layered neural fields to represent collision-free garment surfaces. The key ingredient is a neural untangling projection operator that works directly on the layered neural fields, not on explicit surface representations. Algorithms to resolve object-object interaction are inherently limited by the use of explicit geometric representations, and we show how methods that work directly on neural implicit representations could bring a change of paradigm and open the door to radically different approaches.

## [Geometric Order Learning for Rank Estimation](#)

- Seon-Ho Lee · Nyeong Ho Shin · Chang-Su Kim
- abstract@[open-review](#): A novel approach to rank estimation, called geometric order learning (GOL), is proposed in this paper. First, we construct an embedding space, in which the direction and distance between objects represent order and metric relations between their ranks, by enforcing two geometric constraints: the order constraint compels objects to be sorted according to their ranks, while the metric constraint makes the distance between objects reflect their rank difference. Then, we perform the simple  $\$k\$$  nearest neighbor ( $\$k\$$ -NN) search in the embedding space to estimate the rank of a test object. Moreover, to assess the quality of embedding spaces for rank estimation, we propose a metric called discriminative ratio for ranking (DRR). Extensive experiments on facial age estimation, historical color image (HCI) classification, and aesthetic score regression demonstrate that GOL constructs effective embedding spaces and thus yields excellent rank estimation performances.

## [Practical Adversarial Multivalid Conformal Prediction](#)

- Osbert Bastani · Varun Gupta · Christopher Jung · Georgy Noarov · Ramya Ramalingam · Aaron Roth
- abstract@[open-review](#): We give a simple, generic conformal prediction method for sequential prediction that achieves target empirical coverage guarantees on adversarial data. It is computationally lightweight --- comparable to split conformal prediction --- but does not require having a held-out validation set, and so all data can be used for training models from which to derive a conformal score. Furthermore, it gives stronger than marginal coverage guarantees in two ways. First, it gives threshold-calibrated prediction sets that have correct empirical coverage even conditional on the threshold used to form the prediction set from the conformal score. Second, the user can specify an arbitrary collection of subsets of the feature space --- possibly intersecting --- and the coverage guarantees will also hold conditional on membership in each of these subsets. We call our algorithm MVP, short for MultiValid Prediction. We give both theory and an extensive set of empirical evaluations.

## [Ask4Help: Learning to Leverage an Expert for Embodied Tasks](#)

- Kunal Pratap Singh · Luca Weihs · Alvaro Herrasti · Jonghyun Choi · Aniruddha Kembhavi · Roozbeh Mottaghi
- abstract@[open-review](#): Embodied AI agents continue to become more capable every year with the advent of new models, environments, and benchmarks, but are still far away from being performant and reliable enough to be deployed in real, user-facing, applications. In this paper, we ask: can we bridge this gap by enabling agents to ask for assistance from an expert such as a human being? To this end, we propose the Ask4Help policy that augments agents with the ability to request, and then use expert assistance. Ask4Help policies can be efficiently trained without modifying the original agent's parameters and learn a desirable trade-off between task performance and the amount of requested help, thereby reducing the cost of querying the expert. We evaluate Ask4Help on two different tasks -- object goal navigation and room rearrangement and see substantial improvements in performance using minimal help. On object navigation, an agent that achieves a  $52\%$  success rate is raised to  $86\%$  with  $13\%$  help and for rearrangement, the state-of-the-art model with a  $7\%$  success rate is dramatically improved to  $90.4\%$  using  $39\%$  help. Human trials with Ask4Help demonstrate the efficacy of our approach in practical scenarios.

## [Distributed Methods with Compressed Communication for Solving Variational Inequalities, with Theoretical Guarantees](#)

- Aleksandr Beznosikov · Peter Richtarik · Michael Diskin · Max Ryabinin · Alexander Gasnikov
- abstract@[open-review](#): Variational inequalities in general and saddle point problems in particular are increasingly relevant in machine learning applications, including adversarial learning, GANs, transport and robust optimization. With increasing data and problem sizes necessary to train high performing models across various applications, we need to rely on parallel and distributed computing. However, in distributed training, communication among the compute nodes is a key bottleneck during training, and this problem is exacerbated for high dimensional and over-parameterized models. Due to these considerations, it is important to equip existing methods with strategies that would allow to reduce the volume of transmitted information during training while obtaining a model of comparable quality. In this paper, we present the first theoretically grounded distributed methods for solving variational inequalities and saddle point problems using compressed communication: MASHA1 and MASHA2. Our theory and methods allow for the use of both unbiased (such as Rand $k$ ; MASHA1) and contractive (such as Top $k$ ; MASHA2) compressors. Algorithms supports bidirectional compressions, and also has modifications for stochastic setting with batches and for federated learning with partial participation of clients. We empirically validate our conclusions using two experimental setups: a standard bilinear min-max problem, and large-scale distributed adversarial training of transformers.

## [Bayesian Active Learning with Fully Bayesian Gaussian Processes](#)

- Christoffer Riis · Francisco Antunes · Frederik Høst · Carlos Lima Azevedo · Francisco Pereira
- abstract@[open-review](#): The bias-variance trade-off is a well-known problem in machine learning that only gets more pronounced the less available data there is. In active learning, where labeled data is scarce or difficult to obtain, neglecting this trade-off can cause inefficient and non-optimal querying, leading to unnecessary data labeling. In this paper, we focus on active learning with Gaussian Processes (GPs). We argue that for the GP, the bias-variance trade-off is made by optimization of the two hyperparameters: the length scale and noise-term. Considering that the optimal mode of the joint posterior of the hyperparameters is equivalent to the optimal bias-variance trade-off, we approximate this joint posterior and utilize it to design two new acquisition functions. The first one is a Bayesian variant of Query-by-Committee (B-QBC), and the second is an extension that explicitly minimizes the predictive variance through a Query by Mixture of Gaussian Processes (QB-MGP) formulation. Across six common simulators, we empirically show that B-QBC, on average, achieves the best marginal likelihood, whereas QB-MGP achieves the best predictive performance. We show that incorporating the bias-variance trade-off in the acquisition functions mitigates unnecessary and expensive data labeling.

## [coVariance Neural Networks](#)

- Saurabh Sihag · Gonzalo Mateos · Corey McMillan · Alejandro Ribeiro
- abstract@[open-review](#): Graph neural networks (GNN) are an effective framework that exploit inter-relationships within graph-structured data for learning. Principal component analysis (PCA) involves the projection of data on the eigenspace of the covariance matrix and draws similarities with the graph convolutional filters in GNNs. Motivated by this observation, we study a GNN architecture, called coVariance neural network (VNN), that operates on sample covariance matrices as graphs. We theoretically establish the stability of VNNs to perturbations in the covariance matrix, thus, implying an advantage over standard PCA-based data analysis approaches that are prone to instability due to principal components associated with close eigenvalues. Our experiments on real-world datasets validate our theoretical results and show that VNN performance is indeed more stable than PCA-based statistical approaches. Moreover, our experiments on multi-resolution datasets also demonstrate that VNNs are amenable to transferability of performance over covariance matrices of different dimensions; a feature that is infeasible for PCA-based approaches.

## [Rethinking Lipschitz Neural Networks for Certified L-infinity Robustness](#)

- Bohang Zhang · Du Jiang · Di He · Liwei Wang
- abstract@[open-review](#): Designing neural networks with bounded Lipschitz constant is a promising way to obtain certifiably robust classifiers against adversarial examples. However, the relevant progress for the important  $\ell_\infty$  perturbation setting is rather limited, and a principled understanding of how to design expressive  $\ell_\infty$  Lipschitz networks is still lacking. In this paper, we bridge the gap by studying certified  $\ell_\infty$  robustness from a novel perspective of representing Boolean functions. We derive two fundamental impossibility results that hold for any standard Lipschitz network: one for robust classification on finite datasets, and the other for Lipschitz function approximation. These results identify that networks built upon norm-bounded affine layers and Lipschitz activations intrinsically lose expressive power even in the two-dimensional case, and shed light on how recently

proposed Lipschitz networks (e.g., GroupSort and  $\ell_\infty$ -distance nets) bypass these impossibilities by leveraging order statistic functions. Finally, based on these insights, we develop a unified Lipschitz network that generalizes prior works, and design a practical version that can be efficiently trained (making certified robust training free). Extensive experiments show that our approach is scalable, efficient, and consistently yields better certified robustness across multiple datasets and perturbation radii than prior Lipschitz networks.

## [Autoregressive Search Engines: Generating Substrings as Document Identifiers](#)

- Michele Bevilacqua · Giuseppe Ottaviano · Patrick Lewis · Scott Yih · Sebastian Riedel · Fabio Petroni
- abstract@[open-review](#): Knowledge-intensive language tasks require NLP systems to both provide the correct answer and retrieve supporting evidence for it in a given corpus. Autoregressive language models are emerging as the de-facto standard for generating answers, with newer and more powerful systems emerging at an astonishing pace. In this paper we argue that all this (and future) progress can be directly applied to the retrieval problem with minimal intervention to the models' architecture. Previous work has explored ways to partition the search space into hierarchical structures and retrieve documents by autoregressively generating their unique identifier. In this work we propose an alternative that doesn't force any structure in the search space: using all ngrams in a passage as its possible identifiers. This setup allows us to use an autoregressive model to generate and score distinctive ngrams, that are then mapped to full passages through an efficient data structure. Empirically, we show this not only outperforms prior autoregressive approaches but also leads to an average improvement of at least 10 points over more established retrieval solutions for passage-level retrieval on the KILT benchmark, establishing new state-of-the-art downstream performance on some datasets, while using a considerably lighter memory footprint than competing systems. Code available in the supplementary materials. Pre-trained models will be made available.

## [MAgNet: Mesh Agnostic Neural PDE Solver](#)

- Oussama Boussif · Yoshua Bengio · Loubna Benabbou · Dan Assouline
- abstract@[open-review](#): The computational complexity of classical numerical methods for solving Partial Differential Equations (PDE) scales significantly as the resolution increases. As an important example, climate predictions require fine spatio-temporal resolutions to resolve all turbulent scales in the fluid simulations. This makes the task of accurately resolving these scales computationally out of reach even with modern supercomputers. As a result, current numerical modelers solve PDEs on grids that are too coarse (3km to 200km on each side), which hinders the accuracy and usefulness of the predictions. In this paper, we leverage the recent advances in Implicit Neural Representations (INR) to design a novel architecture that predicts the spatially continuous solution of a PDE given a spatial position query. By augmenting coordinate-based architectures with Graph Neural Networks (GNN), we enable zero-shot generalization to new non-uniform meshes and long-term predictions up to 250 frames ahead that are physically consistent. Our Mesh Agnostic Neural PDE Solver (MAgNet) is able to make accurate predictions across a variety of PDE simulation datasets and compares favorably with existing baselines. Moreover, our model generalizes well to different meshes and resolutions up to four times those trained on.

## [On Embeddings for Numerical Features in Tabular Deep Learning](#)

- Yury Gorishniy · Ivan Rubachev · Artem Babenko
- abstract@[open-review](#): Recently, Transformer-like deep architectures have shown strong performance on tabular data problems. Unlike traditional models, e.g., MLP, these architectures map scalar values of numerical features to high-dimensional embeddings before mixing them in the main backbone. In this work, we argue that embeddings for numerical features are an underexplored degree of freedom in tabular DL, which allows constructing more powerful DL models and competing with gradient boosted decision trees (GBDT) on some GBDT-friendly benchmarks (that is, where GBDT outperforms conventional DL models). We start by describing two conceptually different approaches to building embedding modules: the first one is based on a piecewise linear encoding of scalar values, and the second one utilizes periodic activations. Then, we empirically demonstrate that these two approaches can lead to significant performance boosts compared to the embeddings based on conventional blocks such as linear layers and ReLU activations. Importantly, we also show that embedding numerical features is beneficial for many backbones, not only for Transformers. Specifically, after proper embeddings, simple MLP-like models can perform on par with the attention-based architectures. Overall, we highlight embeddings for numerical features as an important design aspect with good potential for further improvements in tabular DL.

## [BadPrompt: Backdoor Attacks on Continuous Prompts](#)

- Xiangrui Cai · Haidong Xu · Sihan Xu · Ying ZHANG · Yuan xiaojie
- abstract@[open-review](#): The prompt-based learning paradigm has gained much research attention recently. It has achieved state-of-the-art performance on several NLP tasks, especially in the few-shot scenarios. While steering the downstream tasks, few works have been reported to investigate the security problems of the prompt-based models. In this paper, we conduct the first study on the vulnerability of the continuous prompt learning algorithm to backdoor attacks. We observe that the few-shot scenarios have posed a great challenge to backdoor attacks on the prompt-based models, limiting the usability of existing NLP backdoor methods. To address this challenge, we propose BadPrompt, a lightweight and task-adaptive algorithm, to backdoor attack continuous prompts. Specially, BadPrompt first generates candidate triggers which are indicative for predicting the targeted label and dissimilar to the samples of the non-targeted labels. Then, it automatically selects the most effective and invisible trigger for each sample with an adaptive trigger optimization algorithm. We evaluate the performance of BadPrompt on five datasets and two continuous prompt models. The results exhibit the abilities of BadPrompt to effectively attack continuous prompts while maintaining high performance on the clean test sets, outperforming the baseline models by a large margin. The source code of BadPrompt is publicly available.

## [Simulation-guided Beam Search for Neural Combinatorial Optimization](#)

- Jinho Choo · Yeong-Dae Kwon · Jihoon Kim · Jeongwoo Jae · AndrÃ© Hottung · Kevin Tierney · Youngjune Gwon
- abstract@[open-review](#): Neural approaches for combinatorial optimization (CO) equip a learning mechanism to discover powerful heuristics for solving complex real-world problems. While neural approaches capable of high-quality solutions in a single shot are emerging, state-of-the-art approaches are often unable to take full advantage of the solving time available to them. In contrast, hand-crafted heuristics perform highly effective search well and exploit the computation time given to them, but contain heuristics that are difficult to adapt to a dataset being solved. With the goal of providing a powerful search procedure to neural CO approaches, we propose simulation-guided beam search (SGBS), which examines candidate solutions within a fixed-width tree search that both a neural net-learned policy and a simulation (rollout) identify as promising. We further hybridize SGBS with efficient active search (EAS), where SGBS enhances the quality of solutions backpropagated in EAS, and EAS improves the quality of the policy used in SGBS. We evaluate our methods on well-known CO benchmarks and show that SGBS significantly improves the quality of the solutions found under reasonable runtime assumptions.

## [Multi-Scale Adaptive Network for Single Image Denoising](#)

- Yuanbiao Gou · Peng Hu · Jiancheng Lv · Joey Tianyi Zhou · Xi Peng
- abstract@[open-review](#): Multi-scale architectures have shown effectiveness in a variety of tasks thanks to appealing cross-scale complementarity. However, existing methods treat different scale features equally without considering their scale-specific characteristics, i.e., the within-scale characteristics are ignored. In this paper, we reveal this missing piece for multi-scale architecture design and accordingly propose a novel Multi-Scale Adaptive Network (MSANet) for single image denoising. Specifically, MSANet simultaneously embraces the within-scale characteristics and the cross-scale complementarity thanks to three novel neural blocks, adaptive feature block (AFeB), adaptive multi-scale block (AMB), and adaptive fusion block (AFuB). In brief, AFeB is designed to adaptively select details and filter noises, which is highly expected for fine-grained features. AMB could

enlarge the receptive field and aggregate the multi-scale information, which is designed to satisfy the demands of both fine- and coarse-grained features. AFuB devotes to adaptively sampling and transferring the features from one scale to another scale, which is used to fuse the features with varying characteristics from coarse to fine. Extensive experiments on both three real and six synthetic noisy image datasets show the superiority of MSANet compared with 12 methods.

## [An \$\alpha\$ -regret analysis of Adversarial Bilateral Trade](#)

- Yossi Azar · Amos Fiat · Federico Fusco
- abstract@[open-review](#): We study sequential bilateral trade where sellers and buyers valuations are completely arbitrary (i.e., determined by an adversary). Sellers and buyers are strategic agents with private valuations for the good and the goal is to design a mechanism that maximizes efficiency (or gain from trade) while being incentive compatible, individually rational and budget balanced. In this paper we consider gain from trade which is harder to approximate than social welfare. We consider a variety of feedback scenarios and distinguish the cases where the mechanism posts one price and when it can post different prices for buyer and seller. We show several surprising results about the separation between the different scenarios. In particular we show that (a) it is impossible to achieve sublinear  $\alpha$ -regret for any  $\alpha < 2$ , (b) but with full feedback sublinear  $2$ -regret is achievable (c) with a single price and partial feedback one cannot get sublinear  $\alpha$  regret for any constant  $\alpha$  (d) nevertheless, posting two prices even with one bit feedback achieves sublinear  $2$ -regret, and (e) there is a provable separation in the  $2$ -regret bounds between full and partial feedback.

## [Linear-Time Gaussian Processes Using Binary Tree Kernels](#)

- Michael Cohen · Samuel Daulton · Michael A Osborne
- abstract@[open-review](#): Gaussian processes (GPs) produce good probabilistic models of functions, but most GP kernels require  $O((n+m)n^2)$  time, where  $n$  is the number of data points and  $m$  the number of predictive locations. We present a new kernel that allows for Gaussian process regression in  $O((n+m)\log(n+m))$  time. Our "binary tree" kernel places all data points on the leaves of a binary tree, with the kernel depending only on the depth of the deepest common ancestor. We can store the resulting kernel matrix in  $O(n)$  space in  $O(n \log n)$  time, as a sum of sparse rank-one matrices, and approximately invert the kernel matrix in  $O(n)$  time. Sparse GP methods also offer linear run time, but they predict less well than higher dimensional kernels. On a classic suite of regression tasks, we compare our kernel against Matérn, sparse, and sparse variational kernels. The binary tree GP assigns the highest likelihood to the test data on a plurality of datasets, usually achieves lower mean squared error than the sparse methods, and often ties or beats the Matérn GP. On large datasets, the binary tree GP is fastest, and much faster than a Matérn GP.

## [Differentially Private Model Compression](#)

- Fatemeh Sadat Mireshghallah · Arturs Backurs · Huseyin A. Inan · Lukas Wutschitz · Janardhan Kulkarni
- abstract@[open-review](#): Recent papers have shown that large pre-trained language models (LLMs) such as BERT, GPT-2 can be fine-tuned on private data to achieve performance comparable to non-private models for many downstream Natural Language Processing (NLP) tasks while simultaneously guaranteeing differential privacy. The inference cost of these models -- which consist of hundreds of millions of parameters -- however, can be prohibitively large. Hence, often in practice, LLMs are compressed before they are deployed in specific applications. In this paper, we initiate the study of differentially private model compression and propose frameworks for achieving 50% sparsity levels while maintaining nearly full performance. We demonstrate these ideas on standard GLUE benchmarks using BERT models, setting benchmarks for future research on this topic.

## [REVIVE: Regional Visual Representation Matters in Knowledge-Based Visual Question Answering](#)

- Yuanze Lin · Yujia Xie · Dongdong Chen · Yichong Xu · Chenguang Zhu · Lu Yuan
- abstract@[open-review](#): This paper revisits visual representation in knowledge-based visual question answering (VQA) and demonstrates that using regional information in a better way can significantly improve the performance. While visual representation is extensively studied in traditional VQA, it is under-explored in knowledge-based VQA even though these two tasks share the common spirit, i.e., rely on visual input to answer the question. Specifically, we observe in most state-of-the-art knowledge-based VQA methods: 1) visual features are extracted either from the whole image or in a sliding window manner for retrieving knowledge, and the important relationship within/among object regions is neglected; 2) visual features are not well utilized in the final answering model, which is counter-intuitive to some extent. Based on these observations, we propose a new knowledge-based VQA method REVIVE, which tries to utilize the explicit information of object regions not only in the knowledge retrieval stage but also in the answering model. The key motivation is that object regions and inherent relationship are important for knowledge-based VQA. We perform extensive experiments on the standard OK-VQA dataset and achieve new state-of-the-art performance, i.e., 58.0 accuracy, surpassing previous state-of-the-art method by a large margin (+3.6%). We also conduct detailed analysis and show the necessity of regional information in different framework components for knowledge-based VQA.

## [A gradient estimator via L1-randomization for online zero-order optimization with two point feedback](#)

- Arya Akhavan · Evgenii Chzhen · Massimiliano Pontil · Alexandre Tsybakov
- abstract@[open-review](#): This work studies online zero-order optimization of convex and Lipschitz functions. We present a novel gradient estimator based on two function evaluation and randomization on the  $\ell_1$ -sphere. Considering different geometries of feasible sets and Lipschitz assumptions we analyse online mirror descent algorithm with our estimator in place of the usual gradient. We consider two types of assumptions on the noise of the zero-order oracle: canceling noise and adversarial noise. We provide an anytime and completely data-driven algorithm, which is adaptive to all parameters of the problem. In the case of canceling noise that was previously studied in the literature, our guarantees are either comparable or better than state-of-the-art bounds obtained by [duchi2015](#) and [Shamir17](#) for non-adaptive algorithms. Our analysis is based on deriving a new Poincaré type inequality for the uniform measure on the  $\ell_1$ -sphere with explicit constants, which may be of independent interest.

## [A Quantitative Geometric Approach to Neural Network Smoothness](#)

- Zi Wang · Gautam Prakriya · Somesh Jha
- abstract@[open-review](#): Fast and precise Lipschitz constant estimation of neural networks is an important task for deep learning. Researchers have recently found an intrinsic trade-off between the accuracy and smoothness of neural networks, so training a network with a loose Lipschitz constant estimation imposes a strong regularization, and can hurt the model accuracy significantly. In this work, we provide a unified theoretical framework, a quantitative geometric approach, to address the Lipschitz constant estimation. By adopting this framework, we can immediately obtain several theoretical results, including the computational hardness of Lipschitz constant estimation and its approximability. We implement the algorithms induced from this quantitative geometric approach, which are based on semidefinite programming (SDP). Our empirical evaluation demonstrates that they are more scalable and precise than existing tools on Lipschitz constant estimation for  $\ell_\infty$ -perturbations. Furthermore, we also show their intricate relations with other recent SDP-based techniques, both theoretically and empirically. We believe that this unified quantitative geometric perspective can bring new insights and theoretical tools to the investigation of neural-network smoothness and robustness.

## [Sound and Complete Verification of Polynomial Networks](#)

- Elias Abad Rocamora · Mehmet Fatih Sahin · Fanghui Liu · Grigoris Chrysos · Volkan Cevher

- abstract@[open-review](#): Polynomial Networks (PNs) have demonstrated promising performance on face and image recognition recently. However, robustness of PNs is unclear and thus obtaining certificates becomes imperative for enabling their adoption in real-world applications. Existing verification algorithms on ReLU neural networks (NNs) based on branch and bound (BaB) techniques cannot be trivially applied to PN verification. In this work, we devise a new bounding method, equipped with BaB for global convergence guarantees, called VPN. One key insight is that we obtain much tighter bounds than the interval bound propagation baseline. This enables sound and complete PN verification with empirical validation on MNIST, CIFAR10 and STL10 datasets. We believe our method has its own interest to NN verification.

## [Debiased Causal Tree: Heterogeneous Treatment Effects Estimation with Unmeasured Confounding](#)

- Caizhi Tang · Huiyuan Wang · Xinyu Li · Qing Cui · Ya-Lin Zhang · Feng Zhu · Longfei Li · Jun Zhou · Linbo Jiang
- abstract@[open-review](#): Unmeasured confounding poses a significant threat to the validity of causal inference. Despite that various ad hoc methods are developed to remove confounding effects, they are subject to certain fairly strong assumptions. In this work, we consider the estimation of conditional causal effects in the presence of unmeasured confounding using observational data and historical controls. Under an interpretable transportability condition, we prove the partial identifiability of conditional average treatment effect on the treated group (CATT). For tree-based models, a new notion, \emph{confounding entropy}, is proposed to measure the discrepancy introduced by unobserved confounders between the conditional outcome distribution of the treated and control groups. The confounding entropy generalizes conventional confounding bias, and can be estimated effectively using historical controls. We develop a new method, debiased causal tree, whose splitting rule is to minimize the empirical risk regularized by the confounding entropy. Notably, our method integrates current observational data (for empirical risk) and their historical controls (for confounding entropy) harmoniously. We highlight that, debiased causal tree can not only estimate CATT well in the presence of unmeasured confounding, but also is a robust estimator of conditional average treatment effect (CATE) against the imbalance of the treated and control populations when all confounders are observed. An extension of combining multiple debiased causal trees to further reduce biases by gradient boosting is considered. The computational feasibility and statistical power of our method are evidenced by simulations and a study of a credit card balance dataset.

## [SegNeXt: Rethinking Convolutional Attention Design for Semantic Segmentation](#)

- Meng-Hao Guo · Cheng-Ze Lu · Qibin Hou · Zhengning Liu · Ming-Ming Cheng · Shi-min Hu
- abstract@[open-review](#): We present SegNeXt, a simple convolutional network architecture for semantic segmentation. Recent transformer-based models have dominated the field of semantic segmentation due to the efficiency of self-attention in encoding spatial information. In this paper, we show that convolutional attention is still a more efficient and effective way to encode contextual information than the self-attention mechanism in transformers. By re-examining the characteristics owned by successful segmentation models, we discover several key components leading to the performance improvement of segmentation models. This motivates us to design a novel convolutional attention network that uses purely cheap convolutional operations. Without bells and whistles, our SegNeXt significantly improves the performance of previous state-of-the-art methods on popular benchmarks, including ADE20K, Cityscapes, COCO-Stuff, Pascal VOC, Pascal Context, and iSAID. Notably, SegNeXt outperforms EfficientNet-L2 w/ NAS-FPN and achieves 90.6% mIoU on the Pascal VOC 2012 test leaderboard using only 1/10 parameters of it. On average, SegNeXt achieves about 2.0% mIoU improvements compared to the state-of-the-art methods on the ADE20K datasets with the same or fewer computations. Code will be made publicly available.

## [Respecting Transfer Gap in Knowledge Distillation](#)

- Yulei Niu · Long Chen · Hanwang Zhang · Chang Zhou
- abstract@[open-review](#): Knowledge distillation (KD) is essentially a process of transferring a teacher model's behavior, e.g., network response, to a student model. The network response serves as additional supervision to formulate the machine domain, which uses the data collected from the human domain as a transfer set. Traditional KD methods hold an underlying assumption that the data collected in both human domain and machine domain are both independent and identically distributed (IID). We point out that this naive assumption is unrealistic and there is indeed a transfer gap between the two domains. Although the gap offers the student model external knowledge from the machine domain, the imbalanced teacher knowledge would make us incorrectly estimate how much to transfer from teacher to student per sample on the non-IID transfer set. To tackle this challenge, we propose Inverse Probability Weighting Distillation (IPWD) that estimates the propensity of a training sample belonging to the machine domain, and assigns its inverse amount to compensate for under-represented samples. Experiments on CIFAR-100 and ImageNet demonstrate the effectiveness of ours~for both two-stage distillation and one-stage self-distillation.

## [AdaptFormer: Adapting Vision Transformers for Scalable Visual Recognition](#)

- Shoufa Chen · Chongjian GE · Zhan Tong · Jiangliu Wang · Yibing Song · Jue Wang · Ping Luo
- abstract@[open-review](#): Although the pre-trained Vision Transformers (ViTs) achieved great success in computer vision, adapting a ViT to various image and video tasks is challenging because of its heavy computation and storage burdens, where each model needs to be independently and comprehensively fine-tuned to different tasks, limiting its transferability in different domains. To address this challenge, we propose an effective adaptation approach for Transformer, namely AdaptFormer, which can adapt the pre-trained ViTs into many different image and video tasks efficiently. It possesses several benefits more appealing than prior arts. Firstly, AdaptFormer introduces lightweight modules that only adds less than 2% extra parameters to a ViT, while it is able to increase the ViT's transferability without updating its original pre-trained parameters, significantly outperforming the existing 100% fully fine-tuned models on action recognition benchmarks. Secondly, it can be plug-and-play in different Transformers and scalable to many visual tasks. Thirdly, extensive experiments on five image and video datasets show that AdaptFormer largely improves ViTs in the target domains. For example, when updating just 1.5% extra parameters, it achieves about 10% and 19% relative improvement compared to the fully fine-tuned models on Something-Something v2 and HMDB51, respectively. The deliverables are released at anonymous-adaptformer.github.io.

## [Recommender Forest for Efficient Retrieval](#)

- Chao Feng · Wuchao Li · Defu Lian · Zheng Liu · Enhong Chen
- abstract@[open-review](#): Recommender systems (RS) have to select the top-N items from a massive item set. For the sake of efficient recommendation, RS usually represents user and item as latent embeddings, and relies on approximate nearest neighbour search (ANNs) to retrieve the recommendation result. Despite the reduction of running time, the representation learning is independent of ANNs index construction; thus, the two operations can be incompatible, which results in potential loss of recommendation accuracy. To overcome the above problem, we propose the Recommender Forest (a.k.a., RecForest), which jointly learns latent embedding and index for efficient and high-fidelity recommendation. RecForest consists of multiple k-ary trees, each of which is a partition of the item set via hierarchical balanced clustering such that each item is uniquely represented by a path from the root to a leaf. Given such a data structure, an encoder-decoder based routing network is developed: it first encodes the context, i.e., user information, into hidden states; then, leveraging a transformer-based decoder, it identifies the top-N items via beam search. Compared with the existing methods, RecForest brings in the following advantages: 1) the false partition of the boundary items can be effectively alleviated by the use of multiple trees; 2) the routing operation becomes much more accurate thanks to the powerful transformer decoder; 3) the tree parameters are shared across different tree levels, making the index to be extremely memory-efficient. The experimental studies are performed on five popular recommendation datasets: with a significantly simplified training cost, RecForest outperforms competitive baseline approaches in terms of both recommendation accuracy and efficiency.

## [Self-Supervised Image Restoration with Blurry and Noisy Pairs](#)

- Zhilu Zhang · RongJian Xu · Ming Liu · Zifei Yan · Wangmeng Zuo

- abstract@[open-review](#): When taking photos under an environment with insufficient light, the exposure time and the sensor gain usually require to be carefully chosen to obtain images with satisfying visual quality. For example, the images with high ISO usually have inescapable noise, while the long-exposure ones may be blurry due to camera shake or object motion. Existing solutions generally suggest to seek a balance between noise and blur, and learn denoising or deblurring models under either full- or self-supervision. However, the real-world training pairs are difficult to collect, and the self-supervised methods merely rely on blurry or noisy images are limited in performance. In this work, we tackle this problem by jointly leveraging the short-exposure noisy image and the long-exposure blurry image for better image restoration. Such setting is practically feasible due to that short-exposure and long-exposure images can be either acquired by two individual cameras or synthesized by a long burst of images. Moreover, the short-exposure images are hardly blurry, and the long-exposure ones have negligible noise. Their complementarity makes it feasible to learn restoration model in a self-supervised manner. Specifically, the noisy images can be used as the supervision information for deblurring, while the sharp areas in the blurry images can be utilized as the auxiliary supervision information for self-supervised denoising. By learning in a collaborative manner, the deblurring and denoising tasks in our method can benefit each other. Experiments on synthetic and real-world images show the effectiveness and practicality of the proposed method. Source code will be publicly available.

## [Unsupervised Object Detection Pretraining with Joint Object Priors Generation and Detector Learning](#)

- Yizhou Wang Â· Meilin Chen Â· SHIXIANG TANG Â· Feng Zhu Â· Haiyang Yang Â· LEI BAI Â· Rui Zhao Â· Yunfeng Yan Â· Donglian Qi Â· Wanli Ouyang
- abstract@[open-review](#): Unsupervised pretraining methods for object detection aim to learn object discrimination and localization ability from large amounts of images. Typically, recent works design pretext tasks that supervise the detector to predict the defined object priors. They normally leverage heuristic methods to produce object priors, \emph{e.g.,} selective search, which separates the prior generation and detector learning and leads to sub-optimal solutions. In this work, we propose a novel object detection pretraining framework that could generate object priors and learn detectors jointly by generating accurate object priors from the model itself. Specifically, region priors are extracted by attention maps from the encoder, which highlights foregrounds. Instance priors are the selected high-quality output bounding boxes of the detection decoder. By assuming objects as instances in the foreground, we can generate object priors with both region and instance priors. Moreover, our object priors are jointly refined along with the detector optimization. With better object priors as supervision, the model could achieve better detection capability, which in turn promotes the object priors generation. Our method improves the competitive approaches by \textbf{+1.3 AP}, \textbf{+1.7 AP} in 1\% and 10\% COCO low-data regimes object detection. Code shall be released upon acceptance.

## [Learning to Accelerate Partial Differential Equations via Latent Global Evolution](#)

- Tailin Wu Â· Takashi Maruyama Â· Jure Leskovec
- abstract@[open-review](#): Simulating the time evolution of Partial Differential Equations (PDEs) of large-scale systems is crucial in many scientific and engineering domains such as fluid dynamics, weather forecasting and their inverse optimization problems. However, both classical solvers and recent deep learning-based surrogate models are typically extremely computationally intensive, because of their local evolution: they need to update the state of each discretized cell at each time step during inference. Here we develop Latent Evolution of PDEs (LE-PDE), a simple, fast and scalable method to accelerate the simulation and inverse optimization of PDEs. LE-PDE learns a compact, global representation of the system and efficiently evolves it fully in the latent space with learned latent evolution models. LE-PDE achieves speedup by having a much smaller latent dimension to update during long rollout as compared to updating in the input space. We introduce new learning objectives to effectively learn such latent dynamics to ensure long-term stability. We further introduce techniques for speeding-up inverse optimization of boundary conditions for PDEs via backpropagation through time in latent space, and an annealing technique to address the non-differentiability and sparse interaction of boundary conditions. We test our method in a 1D benchmark of nonlinear PDEs, 2D Navier-Stokes flows into turbulent phase and an inverse optimization of boundary conditions in 2D Navier-Stokes flow. Compared to state-of-the-art deep learning-based surrogate models and other strong baselines, we demonstrate up to 128x reduction in the dimensions to update, and up to 15x improvement in speed, while achieving competitive accuracy.

## [On the Effect of Pre-training for Transformer in Different Modality on Offline Reinforcement Learning](#)

- Shiro Takagi
- abstract@[open-review](#): We empirically investigate how pre-training on data of different modalities, such as language and vision, affects fine-tuning of Transformer-based models to Mujoco offline reinforcement learning tasks. Analysis of the internal representation reveals that the pre-trained Transformers acquire largely different representations before and after pre-training, but acquire less information of data in fine-tuning than the randomly initialized one. A closer look at the parameter changes of the pre-trained Transformers reveals that their parameters do not change that much and that the bad performance of the model pre-trained with image data could partially come from large gradients and gradient clipping. To study what information the Transformer pre-trained with language data utilizes, we fine-tune this model with no context provided, finding that the model learns efficiently even without context information. Subsequent follow-up analysis supports the hypothesis that pre-training with language data is likely to make the Transformer get context-like information and utilize it to solve the downstream task.

## [Distinguishing Learning Rules with Brain Machine Interfaces](#)

- Jacob Portes Â· Christian Schmid Â· James M Murray
- abstract@[open-review](#): Despite extensive theoretical work on biologically plausible learning rules, it has been difficult to obtain clear evidence about whether and how such rules are implemented in the brain. We consider biologically plausible supervised- and reinforcement-learning rules and ask whether changes in network activity during learning can be used to distinguish which learning rule is being used. In particular, we note that supervised learning requires a credit-assignment model estimating the mapping from neural activity to behavior and that, in a biological organism, this model will inevitably be an imperfect approximation of the ideal mapping, leading to a bias in the direction of the weight updates relative to the true gradient. Reinforcement learning, on the other hand, requires no credit-assignment model and tends to make weight updates following the true gradient direction. We derive a metric to distinguish between learning rules by observing changes in the network activity during learning, given that the mapping from brain to behavior is known by the experimenter. Because brain-machine interface experiments allow for perfect knowledge of this mapping, we focus on modeling a cursor-control BMI task using recurrent neural networks, showing that learning rules can be distinguished in simulated experiments using only observations that a neuroscience experimenter would plausibly have access to.

## [Dual-discriminative Graph Neural Network for Imbalanced Graph-level Anomaly Detection](#)

- GE ZHANG Â· Zhenyu Yang Â· Jia Wu Â· Jian Yang Â· Shan Xue Â· Hao Peng Â· Jianlin Su Â· Chuan Zhou Â· Quan Z. Sheng Â· Leman Akoglu Â· Charu Aggarwal
- abstract@[open-review](#): Graph-level anomaly detection aims to distinguish anomalous graphs in a graph dataset from normal graphs. Anomalous graphs represent very few but essential patterns in the real world. The anomalous property of a graph may be referable to its anomalous attributes of particular nodes and anomalous substructures referring to a subset of nodes and edges in the graph. In addition, due to the imbalance nature of anomaly problem, the anomalous information will be diluted by normal graphs with overwhelming quantities. Various anomaly notions in the attributes and/or substructures and the imbalance nature together make detecting anomalous graphs a non-trivial task. In this paper, we propose a dual-discriminative graph neural network for graph-level anomaly detection, namely iGAD. Specifically, an anomalous graph attribute-aware graph convolution and an anomalous graph substructure-aware deep Random Walk Kernel (deep RWK) are welded into a graph neural network to achieve a dual-discriminative ability on anomalous attributes and substructures. The deep RWK in iGAD makes up for the deficiency of graph convolution in distinguishing structural information caused by

the simple neighborhood aggregation mechanism. Further, we propose a Point Mutual Information-based loss function to address the imbalance nature of anomaly problem. The loss function enables iGAD to capture the essential correlation between input graphs and their anomalous/normal properties. We evaluate iGAD on four real-world graph datasets. Extensive experiments demonstrate the superiority of iGAD on the graph-level anomaly detection task.

## [Sparse Interaction Additive Networks via Feature Interaction Detection and Sparse Selection](#)

- James Enouen · Yan Liu
- abstract@[open-review](#): There is currently a large gap in performance between the statistically rigorous methods like linear regression or additive splines and the powerful deep methods using neural networks. Previous works attempting to close this gap have failed to fully consider the exponentially growing number of feature combinations which deep networks consider automatically during training. In this work, we develop a tractable selection algorithm to efficiently identify the necessary feature combinations by leveraging techniques in feature interaction detection. Our proposed Sparse Interaction Additive Networks (SIAN) construct a bridge from these simple and interpretable models to a fully connected neural network. SIAN achieves competitive performance against state-of-the-art methods across multiple large-scale tabular datasets and consistently finds an optimal tradeoff between the modeling capacity of neural networks and the generalizability of simpler methods.

## [Efficient and Effective Multi-task Grouping via Meta Learning on Task Combinations](#)

- Xiaozhuang Song · Shun Zheng · Wei Cao · James Yu · Jiang Bian
- abstract@[open-review](#): As a longstanding learning paradigm, multi-task learning has been widely applied into a variety of machine learning applications. Nonetheless, identifying which tasks should be learned together is still a challenging fundamental problem because the possible task combinations grow exponentially with the number of tasks, and existing solutions heavily relying on heuristics may probably lead to ineffective groupings with severe performance degradation. To bridge this gap, we develop a systematic multi-task grouping framework with a new meta-learning problem on task combinations, which is to predict the per-task performance gains of multi-task learning over single-task learning for any combination. Our underlying assumption is that no matter how large the space of task combinations is, the relationships between task combinations and performance gains lie in some low-dimensional manifolds and thus can be learnable. Accordingly, we develop a neural meta learner, MTG-Net, to capture these relationships, and design an active learning strategy to progressively select meta-training samples. In this way, even with limited meta samples, MTG-Net holds the potential to produce reasonable gain estimations on arbitrary task combinations. Extensive experiments on diversified multi-task scenarios demonstrate the efficiency and the effectiveness of our method. Specifically, in a large-scale evaluation with \$27\$ tasks, which produce over one hundred million task combinations, our method almost doubles the performance obtained by the existing best solution given roughly the same computational cost. Data and code are available at <https://anonymous.4open.science/r/MTG-Net-DB77>.

## [CAGroup3D: Class-Aware Grouping for 3D Object Detection on Point Clouds](#)

- Haiyang Wang · lihe Ding · Shaocong Dong · Shaoshuai Shi · Aoxue Li · Jianan Li · Zhenguo Li · Liwei Wang
- abstract@[open-review](#): We present a novel two-stage fully sparse convolutional 3D object detection framework, named CAGroup3D. Our proposed method first generates some high-quality 3D proposals by leveraging the class-aware local group strategy on the object surface voxels with the same semantic predictions, which considers semantic consistency and diverse locality abandoned in previous bottom-up approaches. Then, to recover the features of missed voxels due to incorrect voxel-wise segmentation, we build a fully sparse convolutional RoI pooling module to directly aggregate fine-grained spatial information from backbone for further proposal refinement. It is memory-and-computation efficient and can better encode the geometry-specific features of each 3D proposal. Our model achieves state-of-the-art 3D detection performance with remarkable gains of +3.6% on ScanNet V2 and +2.6% on SUN RGB-D in term of mAP@0.25. Our code and model will be released.

## [Generating Long Videos of Dynamic Scenes](#)

- Tim Brooks · Janne Hellsten · Miika Aittala · Ting-Chun Wang · Timo Aila · Jaakko Lehtinen · Ming-Yu Liu · Alexei Efros · Tero Karras
- abstract@[open-review](#): We present a video generation model that accurately reproduces object motion, changes in camera viewpoint, and new content that arises over time. Existing video generation methods often fail to produce new content as a function of time while maintaining consistencies expected in real environments, such as plausible dynamics and object persistence. A common failure case is for content to never change due to over-reliance on inductive bias to provide temporal consistency, such as a single latent code that dictates content for the entire video. On the other extreme, without long-term consistency, generated videos may morph unrealistically between different scenes. To address these limitations, we prioritize the time axis by redesigning the temporal latent representation and learning long-term consistency from data by training on longer videos. To this end, we leverage a two-phase training strategy, where we separately train using longer videos at a low resolution and shorter videos at a high resolution. To evaluate the capabilities of our model, we introduce two new benchmark datasets with explicit focus on long-term temporal dynamics.

## [CEBaB: Estimating the Causal Effects of Real-World Concepts on NLP Model Behavior](#)

- Eldar D Abraham · Karel D'Oosterlinck · Amir Feder · Yair Gat · Atticus Geiger · Christopher Potts · Roi Reichart · Zhengxuan Wu
- abstract@[open-review](#): The increasing size and complexity of modern ML systems has improved their predictive capabilities but made their behavior harder to explain. Many techniques for model explanation have been developed in response, but we lack clear criteria for assessing these techniques. In this paper, we cast model explanation as the causal inference problem of estimating causal effects of real-world concepts on the output behavior of ML models given actual input data. We introduce CEBaB, a new benchmark dataset for assessing concept-based explanation methods in Natural Language Processing (NLP). CEBaB consists of short restaurant reviews with human-generated counterfactual reviews in which an aspect (food, noise, ambiance, service) of the dining experience was modified. Original and counterfactual reviews are annotated with multiply-validated sentiment ratings at the aspect-level and review-level. The rich structure of CEBaB allows us to go beyond input features to study the effects of abstract, real-world concepts on model behavior. We use CEBaB to compare the quality of a range of concept-based explanation methods covering different assumptions and conceptions of the problem, and we seek to establish natural metrics for comparative assessments of these methods.

## [Adaptively Exploiting d-Separators with Causal Bandits](#)

- Blair Bilodeau · Linbo Wang · Daniel M Roy
- abstract@[open-review](#): Multi-armed bandit problems provide a framework to identify the optimal intervention over a sequence of repeated experiments. Without additional assumptions, minimax optimal performance (measured by cumulative regret) is well-understood. With access to additional observed variables that d-separate the intervention from the outcome (i.e., they are a d-separator), recent "causal bandit" algorithms provably incur less regret. However, in practice it is desirable to be agnostic to whether observed variables are a d-separator. Ideally, an algorithm should be adaptive; that is, perform nearly as well as an algorithm with oracle knowledge of the presence or absence of a d-separator. In this work, we formalize and study this notion of adaptivity, and provide a novel algorithm that simultaneously achieves (a) optimal regret when a d-separator is observed, improving on classical minimax algorithms, and (b) significantly smaller regret than recent causal bandit algorithms when the observed variables are not a d-separator. Crucially, our algorithm does not require any oracle knowledge of whether a d-separator is observed. We also generalize this adaptivity to other conditions, such as the front-door criterion.

## [Private Synthetic Data for Multitask Learning and Marginal Queries](#)

- Giuseppe Vietri · Cedric Archambeau · Sergul Aydore · William Brown · Michael Kearns · Aaron Roth · Ankit Siva · Shuai Tang · Steven Wu
- abstract@[open-review](#): We provide a differentially private algorithm for producing synthetic data simultaneously useful for multiple tasks: marginal queries and multitask machine learning (ML). A key innovation in our algorithm is the ability to directly handle numerical features, in contrast to a number of related prior approaches which require numerical features to be first converted into {high cardinality} categorical features via {a binning strategy}. Higher binning granularity is required for better accuracy, but this negatively impacts scalability. Eliminating the need for binning allows us to produce synthetic data preserving large numbers of statistical queries such as marginals on numerical features, and class conditional linear threshold queries. Preserving the latter means that the fraction of points of each class label above a particular half-space is roughly the same in both the real and synthetic data. This is the property that is needed to train a linear classifier in a multitask setting. Our algorithm also allows us to produce high quality synthetic data for mixed marginal queries, that combine both categorical and numerical features. Our method consistently runs 2-5x faster than the best comparable techniques, and provides significant accuracy improvements in both marginal queries and linear prediction tasks for mixed-type datasets.

## [Peer Prediction for Learning Agents](#)

- Shi Feng · Fang-Yi Yu · Yiling Chen
- abstract@[open-review](#): Peer prediction refers to a collection of mechanisms for eliciting information from human agents when direct verification of the obtained information is unavailable. They are designed to have a game-theoretic equilibrium where everyone reveals their private information truthfully. This result holds under the assumption that agents are Bayesian and they each adopt a fixed strategy across all tasks. Human agents however are observed in many domains to exhibit learning behavior in sequential settings. In this paper, we explore the dynamics of sequential peer prediction mechanisms when participants are learning agents. We first show that the notion of no regret alone for the agents' learning algorithms cannot guarantee convergence to the truthful strategy. We then focus on a family of learning algorithms where strategy updates only depend on agents' cumulative rewards and prove that agents' strategies in the popular Correlated Agreement (CA) mechanism converge to truthful reporting when they use algorithms from this family. This family of algorithms is not necessarily no-regret, but includes several familiar no-regret learning algorithms (e.g multiplicative weight update and Follow the Perturbed Leader) as special cases. Simulation of several algorithms in this family as well as the \$\\epsilon\$-greedy algorithm, which is outside of this family, shows convergence to the truthful strategy in the CA mechanism.

## [Maximizing Revenue under Market Shrinkage and Market Uncertainty](#)

- Maria-Florina Balcan · Siddharth Prasad · Tuomas Sandholm
- abstract@[open-review](#): A shrinking market is a ubiquitous challenge faced by various industries. In this paper we formulate the first formal model of shrinking markets in multi-item settings, and study how mechanism design and machine learning can help preserve revenue in an uncertain, shrinking market. Via a sample-based learning mechanism, we prove the first guarantees on how much revenue can be preserved by truthful multi-item, multi-bidder auctions (for limited supply) when only a random unknown fraction of the population participates in the market. We first present a general reduction that converts any sufficiently rich auction class into a randomized auction robust to market shrinkage. Our main technique is a novel combinatorial construction called a winner diagram that concisely represents all possible executions of an auction on an uncertain set of bidders. Via a probabilistic analysis of winner diagrams, we derive a general possibility result: a sufficiently rich class of auctions always contains an auction that is robust to market shrinkage and market uncertainty. Our result has applications to important practically-constrained settings such as auctions with a limited number of winners. We then show how to efficiently learn an auction that is robust to market shrinkage by leveraging practically-efficient routines for solving the winner determination problem.

## [Flowification: Everything is a normalizing flow](#)

- Bartłomiej Mąjtusiak · Samuel Klein · Tobias Golling · François Fleuret
- abstract@[open-review](#): We develop a method that can be used to turn any multi-layer perceptron or convolutional network into a normalizing flow. In some cases this requires the addition of uncorrelated noise to the model but in the simplest case no additional parameters. The techniques we develop can be applied to a broad range of transformations. Converting standard models to normalizing flows allows the same architectures to be used for a wide range of tasks. Our models also allow existing density estimation techniques to be combined with high performance feature extractors and for the exact likelihood to be calculated. In contrast to standard density estimation techniques that require specific architectures and specialized knowledge, our approach can leverage design knowledge from other domains and is a step closer to the realization of general purpose architectures. We investigate the efficacy of linear and convolutional layers for the task of density estimation on standard datasets. Our results suggest standard layers lack something fundamental that other normalizing flows do not.

## [AttCAT: Explaining Transformers via Attentive Class Activation Tokens](#)

- Yao Qiang · Deng Pan · Chengyin Li · Xin Li · Rhongho Jang · Dongxiao Zhu
- abstract@[open-review](#): Transformers have improved the state-of-the-art in various natural language processing and computer vision tasks. However, the success of the Transformer model has not yet been duly explained. Current explanation techniques, which dissect either the self-attention mechanism or gradient-based attribution, do not necessarily provide a faithful explanation of the inner workings of Transformers due to the following reasons: first, attention weights alone without considering the magnitudes of feature values are not adequate to reveal the self-attention mechanism; second, whereas most Transformer explanation techniques utilize self-attention module, the skip-connection module, contributing a significant portion of information flows in Transformers, has not yet been sufficiently exploited in explanation; third, the gradient-based attribution of individual feature does not incorporate interaction among features in explaining the model's output. In order to tackle the above problems, we propose a novel Transformer explanation technique via attentive class activation tokens, aka, AttCAT, leveraging encoded features, their gradients, and their attention weights to generate a faithful and confident explanation for Transformer's output. Extensive experiments are conducted to demonstrate the superior performance of AttCAT, which generalizes well to different Transformer architectures, evaluation metrics, datasets, and tasks, to the baseline methods.

## [A New Family of Generalization Bounds Using Samplewise Evaluated CMI](#)

- Fredrik Hellström · Giuseppe Durisi
- abstract@[open-review](#): We present a new family of information-theoretic generalization bounds, in which the training loss and the population loss are compared through a jointly convex function. This function is upper-bounded in terms of the disintegrated, samplewise, evaluated conditional mutual information (CMI), an information measure that depends on the losses incurred by the selected hypothesis rather than on the hypothesis itself, as is common in probably approximately correct (PAC)-Bayesian results. We demonstrate the generality of this framework by recovering and extending previously known information-theoretic bounds. Furthermore, using the evaluated CMI, we derive a samplewise, average version of Seeger's PAC-Bayesian bound, where the convex function is the binary KL divergence. In some scenarios, this novel bound results in a tighter characterization of the population loss of deep neural networks than previous bounds. Finally, we derive high-probability versions of some of these average bounds. We demonstrate the unifying nature of the evaluated CMI bounds by showing that they recover average and high-probability generalization bounds for multiclass classification with finite Natarajan dimension.

## [Look where you look! Saliency-guided Q-networks for visual RL tasks](#)

- David Bertoin · Adil Zouitine · Mehdi Zouitine · Emmanuel Rachelson

- abstract@[open-review](#): Deep reinforcement learning policies, despite their outstanding efficiency in simulated visual control tasks, have shown disappointing ability to generalize across disturbances in the input training images. Changes in image statistics or distracting background elements are pitfalls that prevent generalization and real-world applicability of such control policies. We elaborate on the intuition that a good visual policy should be able to identify which pixels are important for its decision, and preserve this identification of important sources of information across images. This implies that training of a policy with small generalization gap should focus on such important pixels and ignore the others. This leads to the introduction of saliency-guided Q-networks (SGQN), a generic method for visual reinforcement learning, that is compatible with any value function learning method. SGQN vastly improves the generalization capability of Soft Actor-Critic agents and outperforms existing state-of-the-art methods on the Deepmind Control Generalization benchmark, setting a new reference in terms of training efficiency, generalization gap, and policy interpretability.

## [Continuous MDP Homomorphisms and Homomorphic Policy Gradient](#)

- Sahand Rezaei-Shoshtari · Rosie Zhao · Prakash Panangaden · David Meger · Doina Precup
- abstract@[open-review](#): Abstraction has been widely studied as a way to improve the efficiency and generalization of reinforcement learning algorithms. In this paper, we study abstraction in the continuous-control setting. We extend the definition of MDP homomorphisms to encompass continuous actions in continuous state spaces. We derive a policy gradient theorem on the abstract MDP, which allows us to leverage approximate symmetries of the environment for policy optimization. Based on this theorem, we propose an actor-critic algorithm that is able to learn the policy and the MDP homomorphism map simultaneously, using the lax bisimulation metric. We demonstrate the effectiveness of our method on benchmark tasks in the DeepMind Control Suite. Our method's ability to utilize MDP homomorphisms for representation learning leads to improved performance when learning from pixel observations.

## [Exploring Non-Monotonic Latent Alignments for Non-Autoregressive Machine Translation](#)

- Chenze Shao · Yang Feng
- abstract@[open-review](#): Non-autoregressive translation (NAT) models are typically trained with the cross-entropy loss, which forces the model outputs to be aligned verbatim with the target sentence and will highly penalize small shifts in word positions. Latent alignment models relax the explicit alignment by marginalizing out all monotonic latent alignments with the CTC loss. However, they cannot handle non-monotonic alignments, which is non-negligible as there is typically global word reordering in machine translation. In this work, we explore non-monotonic latent alignments for NAT. We extend the alignment space to non-monotonic alignments to allow for the global word reordering and further consider all alignments that overlap with the target sentence. We non-monotonically match the alignments to the target sentence and train the latent alignment model to maximize the F1-score of non-monotonic matching. Extensive experiments on major WMT benchmarks show that our method substantially improves the translation performance and achieves comparable performance to the autoregressive Transformer with only one-iteration parallel decoding.

## [Meta-Auto-Decoder for Solving Parametric Partial Differential Equations](#)

- Xiang Huang · Zhanhong Ye · Hongsheng Liu · Shi Ji · Zidong Wang · Kang Yang · Yang Li · Min Wang · Haotian CHU · Fan Yu · Bei Hua · Lei Chen · Bin Dong
- abstract@[open-review](#): Many important problems in science and engineering require solving the so-called parametric partial differential equations (PDEs), i.e., PDEs with different physical parameters, boundary conditions, shapes of computation domains, etc. Recently, building learning-based numerical solvers for parametric PDEs has become an emerging new field. One category of methods such as the Deep Galerkin Method (DGM) and Physics-Informed Neural Networks (PINNs) aim to approximate the solution of the PDEs. They are typically unsupervised and mesh-free, but require going through the time-consuming network training process from scratch for each set of parameters of the PDE. Another category of methods such as Fourier Neural Operator (FNO) and Deep Operator Network (DeepONet) try to approximate the solution mapping directly. Being fast with only one forward inference for each PDE parameter without retraining, they often require a large corpus of paired input-output observations drawn from numerical simulations, and most of them need a predefined mesh as well. In this paper, we propose Meta-Auto-Decoder (MAD), a mesh-free and unsupervised deep learning method that enables the pre-trained model to be quickly adapted to equation instances by implicitly encoding (possibly heterogenous) PDE parameters as latent vectors. The proposed method MAD can be interpreted by manifold learning in infinite-dimensional spaces, granting it a geometric insight. Extensive numerical experiments show that the MAD method exhibits faster convergence speed without losing accuracy than other deep learning-based methods.

## [Theoretical analysis of deep neural networks for temporally dependent observations](#)

- Mingliang Ma · Abolfazl Safikhani
- abstract@[open-review](#): Deep neural networks are powerful tools to model observations over time with non-linear patterns. Despite the widespread use of neural networks in such settings, most theoretical developments of deep neural networks are under the assumption of independent observations, and theoretical results for temporally dependent observations are scarce. To bridge this gap, we study theoretical properties of deep neural networks on modeling non-linear time series data. Specifically, non-asymptotic bounds for prediction error of (sparse) feed-forward neural network with ReLU activation function is established under mixing-type assumptions. These assumptions are mild such that they include a wide range of time series models including auto-regressive models. Compared to independent observations, established convergence rates have additional logarithmic factors to compensate for additional complexity due to dependence among data points. The theoretical results are supported via various numerical simulation settings as well as an application to a macroeconomic data set.

## [Federated Learning from Pre-Trained Models: A Contrastive Learning Approach](#)

- Yue Tan · Guodong Long · Jie Ma · LU LIU · Tianyi Zhou · Jing Jiang
- abstract@[open-review](#): Federated Learning (FL) is a machine learning paradigm that allows decentralized clients to learn collaboratively without sharing their private data. However, excessive computation and communication demands pose challenges to current FL frameworks, especially when training large-scale models. To prevent these issues from hindering the deployment of FL systems, we propose a lightweight framework where clients jointly learn to fuse the representations generated by multiple fixed pre-trained models rather than training a large-scale model from scratch. This leads us to a more practical FL problem by considering how to capture more client-specific and class-relevant information from the pre-trained models and jointly improve each client's ability to exploit those off-the-shelf models. Here, we design a Federated Prototype-wise Contrastive Learning (FedPCL) approach which shares knowledge across clients through their class prototypes and builds client-specific representations in a prototype-wise contrastive manner. Sharing prototypes rather than learnable model parameters allows each client to fuse the representations in a personalized way while keeping the shared knowledge in a compact form for efficient communication. We perform a thorough evaluation of the proposed FedPCL in the lightweight framework, measuring and visualizing its ability to fuse various pre-trained models on popular FL datasets.

## [Online Reinforcement Learning for Mixed Policy Scopes](#)

- Junzhe Zhang · Elias Bareinboim
- abstract@[open-review](#): Combination therapy refers to the use of multiple treatments -- such as surgery, medication, and behavioral therapy - to cure a single disease, and has become a cornerstone for treating various conditions including cancer, HIV, and depression. All possible combinations of treatments lead to a collection of treatment regimens (i.e., policies) with mixed scopes, or what physicians could observe and which actions they should take depending on the context. In this paper, we investigate the online reinforcement learning setting for optimizing the policy space with mixed scopes. In

particular, we develop novel online algorithms that achieve sublinear regret compared to an optimal agent deployed in the environment. The regret bound has a dependency on the maximal cardinality of the induced state-action space associated with mixed scopes. We further introduce a canonical representation for an arbitrary subset of interventional distributions given a causal diagram, which leads to a non-trivial, minimal representation of the model parameters.

## [Sparse Structure Search for Parameter-Efficient Tuning](#)

- Shengding Hu · Zhen Zhang · Ning Ding · Yadao Wang · Yasheng Wang · Zhiyuan Liu · Maosong Sun
- abstract@[open-review](#): Adapting large pre-trained models (PTMs) through fine-tuning imposes prohibitive computational and storage burdens. Recent studies of parameter-efficient tuning (PET) find that only optimizing a small portion of parameters conditioned on PTMs could yield on-par performance compared to conventional fine-tuning. Generally, PET methods exquisitely design parameter-efficient modules (PET modules) which could be applied to arbitrary fine-grained positions inside PTMs. However, the effectiveness of these fine-grained positions largely relies on sophisticated manual designation, thereby usually producing sub-optimal results. In contrast to the manual designation, we explore constructing PET modules in an automatic manner. We automatically Search for the Sparse Structure of Parameter Efficient Tuning (S\$^3\$PET). Based on a unified framework of various PET methods, S\$^3\$PET conducts the differentiable PET structure search through bi-level optimization and proposes shifted global sigmoid method to explicitly control the number of trainable parameters. Extensive experiments show that S\$^3\$PET surpasses manual and random structures with less trainable parameters. The searched structures preserve more than 99% fine-tuning performance with 0.01% trainable parameters. Moreover, the advantage of S\$^3\$PET is amplified with extremely low trainable parameters budgets (0.0009% \$\sim\$ 0.01%). The searched structures are transferable and explainable, providing suggestions and guidance for the future design of PET methods.

## [ConvMAE: Masked Convolution Meets Masked Autoencoders](#)

- Peng Gao · Teli Ma · Hongsheng Li · Ziyi Lin · Jifeng Dai · Yu Qiao
- abstract@[open-review](#): Vision Transformers (ViT) become widely-adopted architectures for various vision tasks. Masked auto-encoding for feature pretraining and multi-scale hybrid convolution-transformer architectures can further unleash the potentials of ViT, leading to state-of-the-art performances on image classification, detection and semantic segmentation. In this paper, our ConvMAE framework demonstrates that multi-scale hybrid convolution-transformer can learn more discriminative representations via the mask auto-encoding scheme. However, directly using the original masking strategy leads to the heavy computational cost and pretraining-finetuning discrepancy. To tackle the issue, we adopt the masked convolution to prevent information leakage in the convolution blocks. A simple block-wise masking strategy is proposed to ensure computational efficiency. We also propose to more directly supervise the multi-scale features of the encoder to boost multi-scale features. Based on our pretrained ConvMAE models, ConvMAE-Base improves ImageNet-1K finetuning accuracy by 1.4% compared with MAE-Base. On object detection, ConvMAE-Base finetuned for only 25 epochs surpasses MAE-Base fined-tuned for 100 epochs by 2.9% box AP and 2.2% mask AP respectively.

## [On the Spectral Bias of Convolutional Neural Tangent and Gaussian Process Kernels](#)

- Amnon Geifman · Meirav Galun · David Jacobs · Basri Ronen
- abstract@[open-review](#): We study the properties of various over-parameterized convolutional neural architectures through their respective Gaussian process and neural tangent kernels. We prove that, with normalized multi-channel input and ReLU activation, the eigenfunctions of these kernels with the uniform measure are formed by products of spherical harmonics, defined over the channels of the different pixels. We next use hierarchical factorizable kernels to bound their respective eigenvalues. We show that the eigenvalues decay polynomially, quantify the rate of decay, and derive measures that reflect the composition of hierarchical features in these networks. Our theory provides a concrete quantitative characterization of the role of locality and hierarchy in the inductive bias of over-parameterized convolutional network architectures.

## [Discovering Design Concepts for CAD Sketches](#)

- Yuezhi Yang · Hao Pan
- abstract@[open-review](#): Sketch design concepts are recurring patterns found in parametric CAD sketches. Though rarely explicitly formalized by the CAD designers, these concepts are implicitly used in design for modularity and regularity. In this paper, we propose a learning based approach that discovers the modular concepts by induction over raw sketches. We propose the dual implicit-explicit representation of concept structures that allows implicit detection and explicit generation, and the separation of structure generation and parameter instantiation for parameterized concept generation, to learn modular concepts by end-to-end training. We demonstrate the design concept learning on a large scale CAD sketch dataset and show its applications for design intent interpretation and auto-completion.

## [SAViT: Structure-Aware Vision Transformer Pruning via Collaborative Optimization](#)

- Chuanyang Zheng · zheyang li · Kai Zhang · Zhi Yang · Wenming Tan · Jun Xiao · Ye Ren · Shiliang Pu
- abstract@[open-review](#): Vision Transformers (ViTs) yield impressive performance across various vision tasks. However, heavy computation and memory footprint make them inaccessible for edge devices. Previous works apply importance criteria determined independently by each individual component to prune ViTs. Considering that heterogeneous components in ViTs play distinct roles, these approaches lead to suboptimal performance. In this paper, we introduce joint importance, which integrates essential structural-aware interactions between components for the first time, to perform collaborative pruning. Based on the theoretical analysis, we construct a Taylor-based approximation to evaluate the joint importance. This guides pruning toward a more balanced reduction across all components. To further reduce the algorithm complexity, we incorporate the interactions into the optimization function under some mild assumptions. Moreover, the proposed method can be seamlessly applied to various tasks including object detection. Extensive experiments demonstrate the effectiveness of our method. Notably, the proposed approach outperforms the existing state-of-the-art approaches on ImageNet, increasing accuracy by 0.7% over the DeiT-Base baseline while saving 50% FLOPs. On COCO, we are the first to show that 70% FLOPs of FasterRCNN with ViT backbone can be removed with only 0.3% mAP drop. Code will be made available soon.

## [Gradient Descent: The Ultimate Optimizer](#)

- Kartik Chandra · Audrey Xie · Jonathan Ragan-Kelley · ERIK MEIJER
- abstract@[open-review](#): Working with any gradient-based machine learning algorithm involves the tedious task of tuning the optimizer's hyperparameters, such as the step size. Recent work has shown how the step size can itself be "learned" on-line by gradient descent, by manually deriving expressions for "hypergradients" ahead of time. We show how to automatically compute hypergradients with a simple and elegant modification to backpropagation. This allows us to apply the method to other hyperparameters besides the step size, such as the momentum coefficient. We can even recursively apply the method to its own hyperparameters, and so on ad infinitum. As these towers of optimizers grow taller, they become less sensitive to the initial choice of hyperparameters. We present experiments validating this for MLPs, CNNs, and RNNs.

## [Online Deep Equilibrium Learning for Regularization by Denoising](#)

- Jiaming Liu · Xiaojian Xu · Weijie Gan · Shirin Shoushtari · Ulugbek Kamilov

- abstract@[open-review](#): Plug-and-Play Priors (PnP) and Regularization by Denoising (RED) are widely-used frameworks for solving imaging inverse problems by computing fixed-points of operators combining physical measurement models and learned image priors. While traditional PnP/RED formulations have focused on priors specified using image denoisers, there is a growing interest in learning PnP/RED priors that are end-to-end optimal. The recent Deep Equilibrium Models (DEQ) framework has enabled memory-efficient end-to-end learning of PnP/RED priors by implicitly differentiating through the fixed-point equations without storing intermediate activation values. However, the dependence of the computational/memory complexity of the measurement models in PnP/RED on the total number of measurements leaves DEQ impractical for many imaging applications. We propose ODER as a new strategy for improving the efficiency of DEQ through stochastic approximations of the measurement models. We theoretically analyze ODER giving insights into its convergence and ability to approximate the traditional DEQ approach. Our numerical results suggest the potential improvements in training/testing complexity due to ODER on three distinct imaging applications.

## [Exploring Figure-Ground Assignment Mechanism in Perceptual Organization](#)

- Wei Zhai · Yang Cao · Jing Zhang · Zheng-Jun Zha
- abstract@[open-review](#): Perceptual organization is a challenging visual task that aims to perceive and group the individual visual element so that it is easy to understand the meaning of the scene as a whole. Most recent methods building upon advanced Convolutional Neural Network (CNN) come from learning discriminative representation and modeling context hierarchically. However, when the visual appearance difference between foreground and background is obscure, the performance of existing methods degrades significantly due to the visual ambiguity in the discrimination process. In this paper, we argue that the figure-ground assignment mechanism, which conforms to human vision cognitive theory, can be explored to empower CNN to achieve a robust perceptual organization despite visual ambiguity. Specifically, we present a novel Figure-Ground-Aided (FGA) module to learn the configural statistics of the visual scene and leverage it for the reduction of visual ambiguity. Particularly, we demonstrate the benefit of using stronger supervisory signals by teaching (FGA) module to perceive configural cues, i.e., convexity and lower region, that human deem important for the perceptual organization. Furthermore, an Interactive Enhancement Module (IEM) is devised to leverage such configural priors to assist representation learning, thereby achieving robust perception organization with complex visual ambiguities. In addition, a well-founded visual segregation test is designed to validate the capability of the proposed FGA mechanism explicitly. Comprehensive evaluation results demonstrate our proposed FGA mechanism can effectively enhance the capability of perception organization on various baseline models. Nevertheless, the model augmented via our proposed FGA mechanism also outperforms state-of-the-art approaches on four challenging real-world applications. The source code will be made available to the public.

## [Near-Optimal Multi-Agent Learning for Safe Coverage Control](#)

- Manish Prajapat · Matteo Turchetta · Melanie Zeilinger · Andreas Krause
- abstract@[open-review](#): In multi-agent coverage control problems, agents navigate their environment to reach locations that maximize the coverage of some density. In practice, the density is rarely known  $\$textit{a priori}$ , further complicating the original NP-hard problem. Moreover, in many applications, agents cannot visit arbitrary locations due to  $\$textit{a priori}$  unknown safety constraints. In this paper, we aim to efficiently learn the density to approximately solve the coverage problem while preserving the agents' safety. We first propose a conditionally linear submodular coverage function that facilitates theoretical analysis. Utilizing this structure, we develop MacOpt, a novel algorithm that efficiently trades off the exploration-exploitation dilemma due to partial observability, and show that it achieves sublinear regret. Next, we extend results on single-agent safe exploration to our multi-agent setting and propose SafeMac for safe coverage and exploration. We analyze SafeMac and give first of its kind results: near optimal coverage in finite time while provably guaranteeing safety. We extensively evaluate our algorithms on synthetic and real problems, including a bio-diversity monitoring task under safety constraints, where SafeMac outperforms competing methods.

## [SQ Lower Bounds for Learning Single Neurons with Massart Noise](#)

- Ilias Diakonikolas · Daniel Kane · Lisheng Ren · Yuxin Sun
- abstract@[open-review](#): We study the problem of PAC learning a single neuron in the presence of Massart noise. Specifically, for a known activation function  $f: \mathbb{R} \rightarrow \mathbb{R}$ , the learner is given access to labeled examples  $(\mathbf{x}, y) \in \mathbb{R}^d \times \mathbb{R}$ , where the marginal distribution of  $\mathbf{x}$  is arbitrary and the corresponding label  $y$  is a Massart corruption of  $f(\langle \mathbf{x}, \mathbf{w} \rangle)$ . The goal of the learner is to output a hypothesis  $h: \mathbb{R}^d \rightarrow \mathbb{R}$  with small squared loss. For a range of activation functions, including ReLUs, we establish super-polynomial Statistical Query (SQ) lower bounds for this learning problem. In more detail, we prove that no efficient SQ algorithm can approximate the optimal error within any constant factor. Our main technical contribution is a novel SQ-hard construction for learning  $\$pm 1$ -weight Massart halfspaces on the Boolean hypercube that is interesting on its own right.

## [Relational Proxies: Emergent Relationships as Fine-Grained Discriminators](#)

- ABHRA CHAUDHURI · Massimiliano Mancini · Zeynep Akata · Anjan Dutta
- abstract@[open-review](#): Fine-grained categories that largely share the same set of parts cannot be discriminated based on part information alone, as they mostly differ in the way the local parts relate to the overall global structure of the object. We propose Relational Proxies, a novel approach that leverages the relational information between the global and local views of an object for encoding its semantic label. Starting with a rigorous formalization of the notion of distinguishability between fine-grained categories, we prove the necessary and sufficient conditions that a model must satisfy in order to learn the underlying decision boundaries in the fine-grained setting. We design Relational Proxies based on our theoretical findings and evaluate it on six challenging fine-grained benchmark datasets and achieve state-of-the-art results on all of them, surpassing the performance of all existing works with a margin exceeding 4% in some cases. We also experimentally validate our theory on fine-grained distinguishability and obtain consistent results across multiple benchmarks. Code and pre-trained models will be made public upon acceptance.

## [When Privacy Meets Partial Information: A Refined Analysis of Differentially Private Bandits](#)

- Achraf Azize · Debabrota Basu
- abstract@[open-review](#): We study the problem of multi-armed bandits with  $\hat{\mu}$ -global Differential Privacy (DP). First, we prove the minimax and problem-dependent regret lower bounds for stochastic and linear bandits that quantify the hardness of bandits with  $\hat{\mu}$ -global DP. These bounds suggest the existence of two hardness regimes depending on the privacy budget  $\hat{\mu}$ . In the high-privacy regime (small  $\hat{\mu}$ ), the hardness depends on a coupled effect of privacy and partial information about the reward distributions. In the low-privacy regime (large  $\hat{\mu}$ ), bandits with  $\hat{\mu}$ -global DP are not harder than the bandits without privacy. For stochastic bandits, we further propose a generic framework to design a near-optimal  $\hat{\mu}$  global DP extension of an index-based optimistic bandit algorithm. The framework consists of three ingredients: the Laplace mechanism, arm-dependent adaptive episodes, and usage of only the rewards collected in the last episode for computing private statistics. Specifically, we instantiate  $\hat{\mu}$ -global DP extensions of UCB and KL-UCB algorithms, namely AdaP-UCB and AdaP-KLUCB. AdaP-KLUCB is the first algorithm that both satisfies  $\hat{\mu}$ -global DP and yields a regret upper bound that matches the problem-dependent lower bound up to multiplicative constants.

## [Cross-Linked Unified Embedding for cross-modality representation learning](#)

- Xinming Tu · Zhi-Jie Cao · Xia Chenrui · Sara Mostafavi · Ge Gao
- abstract@[open-review](#): Multimodal learning is essential for understanding information in the real world. Jointly learning from multi-modal data enables global integration of both shared and modality-specific information, but current strategies often fail when observations from certain modalities are incomplete/missing for part of the subjects. To learn comprehensive representations based on such modality-incomplete data, we present a semi-

supervised neural network model called CLUE (Cross-Linked Unified Embedding). Extending from multimodal VAEs, CLUE introduces the use of cross-encoders to construct comprehensive representations from modality-incomplete observations. Human cells are tightly regulated across multiple related but distinct modalities such as DNA, RNA, and protein. These modalities jointly define a cell's fate. We benchmark CLUE on multi-modal data from single cell measurements, and show CLUE's superior performance over states-of-the-art on the multi-modal single-cell data integration task, achieving a decisive win in a previous competition. We note that the proposed cross-linked embedding strategy could be readily applied to other popular cross-modality representation learning problems.

## [Large-batch Optimization for Dense Visual Predictions](#)

- Zeyue Xue · Jianming Liang · Guanglu Song · Zhuofan Zong · Liang Chen · Yu Liu · Ping Luo
- abstract@[open-review](#): Training a large-scale deep neural network in a large-scale dataset is challenging and time-consuming. The recent breakthrough of large-batch optimization is a promising way to tackle this challenge. However, although the current advanced algorithms such as LARS and LAMB succeed in classification models, the complicated pipelines of dense visual predictions such as object detection and segmentation still suffer from the heavy performance drop in the large-batch training regime. To address this challenge, we propose a simple yet effective algorithm, named Adaptive Gradient Variance Modulator (AGVM), which can train dense visual predictors with very large batch size, enabling several benefits more appealing than prior arts. Firstly, AGVM can align the gradient variances between different modules in the dense visual predictors, such as backbone, feature pyramid network (FPN), detection, and segmentation heads. We show that training with a large batch size can fail with the gradient variances misaligned among them, which is a phenomenon primarily overlooked in previous work. Secondly, AGVM is a plug-and-play module that generalizes well to many different architectures (e.g., CNNs and Transformers) and different tasks (e.g., object detection, instance segmentation, semantic segmentation, and panoptic segmentation). It is also compatible with different optimizers (e.g., SGD and AdamW). Thirdly, a theoretical analysis of AGVM is provided. Extensive experiments on the COCO and ADE20K datasets demonstrate the superiority of AGVM. For example, AGVM demonstrates more stable generalization performance than prior arts under extremely large batch size (i.e., 10k). AGVM can train Faster R-CNN+ResNet50 in 4 minutes without losing performance. It enables training an object detector with one billion parameters in just 3.5 hours, reducing the training time by 20.9%—, whilst achieving 62.2 mAP on COCO. The deliverables will be released at <https://anonymized-agvm.github.io/>.

## [Adaptive Oracle-Efficient Online Learning](#)

- Guanghui Wang · Zihao Hu · Vidya Muthukumar · Jacob Abernethy
- abstract@[open-review](#): The classical algorithms for online learning and decision-making have the benefit of achieving the optimal performance guarantees, but suffer from computational complexity limitations when implemented at scale. More recent sophisticated techniques, which we refer to as \$\\textit{oracle-efficient}\$ methods, address this problem by dispatching to an \$\\textit{offline optimization oracle}\$ that can search through an exponentially-large (or even infinite) space of decisions and select that which performed the best on any dataset. But despite the benefits of computational feasibility, most oracle-efficient algorithms exhibit one major limitation: while performing well in worst-case settings, they do not adapt well to friendly environments. In this paper we consider two such friendly scenarios, (a) "small-loss" problems and (b) IID data. We provide a new framework for designing follow-the-perturbed-leader algorithms that are oracle-efficient and adapt well to the small-loss environment, under a particular condition which we call \$\\textit{approximability}\$ (which is spiritually related to sufficient conditions provided in (Dud\\'ak et al., 2020)). We identify a series of real-world settings, including online auctions and transductive online classification problems, for which approximability holds. We also extend the algorithm to an IID data setting and establish a "best-of-both-worlds" bound in the oracle-efficient setting.

## [Few-Shot Continual Active Learning by a Robot](#)

- Ali Ayub · Carter Fendley
- abstract@[open-review](#): Most continual learning methods proposed in the literature are focused on task-based continual learning setup. In this setup, a CL model learns a sequence of tasks, one at a time, with all data of the current task labeled and available in an increment, but not of previous or future tasks. This setup, however, is rarely encountered in real-world robotics applications, where a robot might get limited supervision from its users to learn new tasks. Therefore, in this paper, we consider a challenging but realistic continual learning problem, Few-Shot Continual Active Learning (FoCAL), where a CL agent is provided with unlabeled data for a new or a previously learned task in each increment and the agent only has limited labeling budget available. Towards this, we build on the continual learning and active learning literature and develop a framework that can allow a CL agent to continually learn new object classes from a few labeled training examples. Our framework represents each object class using a uniform Gaussian mixture model (GMM) and uses pseudo-rehearsal to mitigate catastrophic forgetting. The framework also uses uncertainty measures on the Gaussian representations of the previously learned classes to find the most informative samples to be labeled in an increment. We evaluate our approach on the CORe-50 dataset and on a real humanoid robot for the object classification task. The results show that our approach not only produces state-of-the-art results on the dataset but also allows a real robot to continually learn unseen objects in a real environment with limited labeling supervision provided by its user.

## [SAPA: Similarity-Aware Point Affiliation for Feature Upsampling](#)

- Hao Lu · Wenze Liu · Zixuan Ye · Hongtao Fu · Yuliang Liu · Zhiguo Cao
- abstract@[open-review](#): We introduce point affiliation into feature upsampling, a notion that describes the affiliation of each upsampled point to a semantic cluster formed by local decoder feature points with semantic similarity. By rethinking point affiliation, we present a generic formulation for generating upsampling kernels. The kernels encourage not only semantic smoothness but also boundary sharpness in the upsampled feature maps. Such properties are particularly useful for some dense prediction tasks such as semantic segmentation. The key idea of our formulation is to generate similarity-aware kernels by comparing the similarity between each encoder feature point and the spatially associated local region of decoder features. In this way, the encoder feature point can function as a cue to inform the semantic cluster of upsampled feature points. To embody the formulation, we further instantiate a lightweight upsampling operator, termed Similarity-Aware Point Affiliation (SAPA), and investigate its variants. SAPA invites consistent performance improvements on a number of dense prediction tasks, including semantic segmentation, object detection, depth estimation, and image matting. Code is available at: <https://github.com/poppinace/sapa>

## [SAPipe: Staleness-Aware Pipeline for Data Parallel DNN Training](#)

- Yangrui Chen · Cong Xie · Meng Ma · Juncheng Gu · Yanghua Peng · Haibin Lin · Chuan Wu · Yibo Zhu
- abstract@[open-review](#): Data parallelism across multiple machines is widely adopted for accelerating distributed deep learning, but it is hard to achieve linear speedup due to the heavy communication. In this paper, we propose SAPipe, a performant system that pushes the training speed of data parallelism to its fullest extent. By introducing partial staleness, the communication overlaps the computation with minimal staleness in SAPipe. To mitigate additional problems incurred by staleness, SAPipe adopts staleness compensation techniques including weight prediction and delay compensation with provably lower error bounds. Additionally, SAPipe presents an algorithm-system co-design with runtime optimization to minimize system overhead for the staleness training pipeline and staleness compensation. We have implemented SAPipe in the BytePS framework, compatible to both TensorFlow and PyTorch. Our experiments show that SAPipe achieves up to 157% speedups over BytePS (non-stale), and outperforms PipeSGD in accuracy by up to 13.7%.

## [Infinite Recommendation Networks: A Data-Centric Approach](#)

- Noveen Sachdeva · Mehak Dhaliwal · Carole-Jean Wu · Julian McAuley

- abstract@[open-review](#): We leverage the Neural Tangent Kernel and its equivalence to training infinitely-wide neural networks to devise \$\text{Infty-AE}\$: an autoencoder with infinitely-wide bottleneck layers. The outcome is a highly expressive yet simplistic recommendation model with a single hyper-parameter and a closed-form solution. Leveraging \$\text{Infty-AE}\$'s simplicity, we also develop Distill-CF for synthesizing tiny, high-fidelity data summaries which distill the most important knowledge from the extremely large and sparse user-item interaction matrix for efficient and accurate subsequent data-usage like model training, inference, architecture search, etc. This takes a data-centric approach to recommendation, where we aim to improve the quality of logged user-feedback data for subsequent modeling, independent of the learning algorithm. We particularly utilize the concept of differentiable Gumbel-sampling to handle the inherent data heterogeneity, sparsity, and semi-structuredness, while being scalable to datasets with hundreds of millions of user-item interactions. Both of our proposed approaches significantly outperform their respective state-of-the-art and when used together, we observe \$96\$-\$110\%\$ of \$\text{Infty-AE}\$'s performance on the full dataset with as little as \$0.1\%\$ of the original dataset size, leading us to explore the counter-intuitive question: Is more data what you need for better recommendation?

## [RepLAI: Self-supervised Representation Learning from Videos of Audible Interactions](#)

- Himangi Mittal · Pedro Morgado · Unnat Jain · Abhinav Gupta
- abstract@[open-review](#): We propose a self-supervised algorithm to learn representations from egocentric video data. Recently, significant efforts have been made to capture humans interacting with their own environments as they go about their daily activities. In result, several large egocentric datasets of interaction-rich multi-modal data have emerged. However, learning representations from videos can be challenging. First, given the uncurated nature of long-form continuous videos, learning effective representations require focusing on moments in time when interactions take place. Second, visual representations of daily activities should be sensitive to changes in the state of the environment. However, current successful multi-modal learning frameworks encourage representation invariance over time. To address these challenges, we leverage audio signals to identify moments of likely interactions which are conducive to better learning. We also propose a novel self-supervised objective that learns from audible state changes caused by interactions. We validate these contributions extensively on two large-scale egocentric datasets, EPIC-Kitchens-100 and the recently released Ego4D, and show improvements on several downstream tasks, including action recognition, long-term action anticipation, and object state change classification.

## [Tackling Overfitting and Silence in Unsupervised Audio-Visual Source Localization](#)

- Shentong Mo · Pedro Morgado
- abstract@[open-review](#): Audio-visual source localization is a challenging task that aims to predict the location of visual sound sources in a video. Since collecting ground-truth annotations of sounding objects can be costly, a plethora of weakly-supervised localization methods that can learn from datasets with no bounding-box annotations have been proposed in recent years, by leveraging the natural co-occurrence of audio and visual signals. Despite significant interest, current evaluation protocols have two major flaws. First, they allow for the use of a fully annotated dataset to perform early stopping, thus significantly increasing the annotation effort required for training. Second, current evaluation metrics assume the presence of sound sources at all times. This is of course an unrealistic assumption, and thus better metrics are necessary to capture the model's performance on (negative) samples with no visible sound sources. To accomplish this, we extend the test set of popular benchmarks, Flickr SoundNet and VGG-Sound Sources, in order to include negative samples, and measure performance using metrics that balance precision and recall. Using the new protocol, we conducted an extensive evaluation of prior methods, and found that most prior works are not capable of identifying negatives and suffer from significant overfitting problems (rely heavily on early stopping for best results). We also propose a new approach for visual sound source localization that addresses both these problems. In particular, we found that, through extreme visual dropout and the use of momentum encoders, the proposed approach combats overfitting effectively, and establishes a new state-of-the-art performance on both Flickr SoundNet and VGG-Sound Source.

## [Sym-NCO: Leveraging Symmetricity for Neural Combinatorial Optimization](#)

- Minsu Kim · Junyoung Park · Jinkyoo Park
- abstract@[open-review](#): Deep reinforcement learning (DRL)-based combinatorial optimization (CO) methods (i.e., DRL-NCO) have shown significant merit over the conventional CO solvers as DRL-NCO is capable of learning CO solvers less relying on problem-specific expert domain knowledge (heuristic method) and supervised labeled data (supervised learning method). This paper presents a novel training scheme, Sym-NCO, which is a regularizer-based training scheme that leverages universal symmetricities in various CO problems and solutions. Leveraging symmetricities such as rotational and reflectional invariance can greatly improve the generalization capability of DRL-NCO because it allows the learned solver to exploit the commonly shared symmetricities in the same CO problem class. Our experimental results verify that our Sym-NCO greatly improves the performance of DRL-NCO methods in four CO tasks, including traveling salesman problem (TSP), capacitated vehicle routing problem (CVRP), prize collecting TSP (PCTSP), and orienteering problem (OP), without utilizing problem-specific expert domain knowledge. Remarkably, Sym-NCO outperformed not only the existing DRL-NCO methods but also a competitive conventional solver, the iterative local search (ILS), in PCTSP at 240\$\times\$ faster speed. Source code will be available after the decision is made.

## [Toward Equation of Motion for Deep Neural Networks: Continuous-time Gradient Descent and Discretization Error Analysis](#)

- Taiki Miyagawa
- abstract@[open-review](#): We derive and solve an ``Equation of Motion'' (EoM) for deep neural networks (DNNs), a differential equation that precisely describes the discrete learning dynamics of DNNs. Differential equations are continuous but have played a prominent role even in the study of discrete optimization (gradient descent (GD) algorithms). However, there still exist gaps between differential equations and the actual learning dynamics of DNNs due to discretization error. In this paper, we start from gradient flow (GF) and derive a counter term that cancels the discretization error between GF and GD. As a result, we obtain EoM, a continuous differential equation that precisely describes the discrete learning dynamics of GD. We also derive discretization error to show to what extent EoM is precise. In addition, we apply EoM to two specific cases: scale- and translation-invariant layers. EoM highlights differences between continuous and discrete GD, indicating the importance of the counter term for a better description of the discrete learning dynamics of GD. Our experimental results support our theoretical findings.

## [RecZilla: Algorithm Selection for Recommender Systems](#)

- Duncan McElfresh · Sujay Khandagale · Jonathan Valverde · John Dickerson · Colin White
- abstract@[open-review](#): While other areas of machine learning have seen more and more automation, designing a high-performing recommender system still requires a high level of human effort. Furthermore, recent work has shown that modern recommender system algorithms do not always improve over well-tuned baselines. A natural follow-up question is, "how do we choose the right algorithm for a new dataset and performance metric?" In this work, we start by giving the first large-scale study of recommender system approaches by comparing 24 algorithms and 100 sets of hyperparameters across 85 datasets and 315 metrics. We find that the best algorithms and hyperparameters are highly dependent on the dataset and performance metric, however, there is also a strong correlation between the performance of each algorithm and various meta-features of the datasets. Motivated by these findings, we create RecZilla, a meta-learning approach to recommender systems that uses a model to predict the best algorithm and hyperparameters for new, unseen datasets. By using far more meta-training data than prior work, RecZilla is able to substantially reduce the level of human involvement when faced with a new recommender system application. We not only release our code and pretrained RecZilla models, but also all of our raw experimental results, so that practitioners can train a RecZilla model for their desired performance metric: <https://anonymous.4open.science/r/anon-reczilla-51FC>.

## [A Contrastive Framework for Neural Text Generation](#)

- Yixuan Su · Tian Lan · Yan Wang · Dani Yogatama · Lingpeng Kong · Nigel Collier
- abstract@[open-review](#): Text generation is of great importance to many natural language processing applications. However, maximization-based decoding methods (e.g., beam search) of neural language models often lead to degenerate solutions---the generated text is unnatural and contains undesirable repetitions. Existing approaches introduce stochasticity via sampling or modify training objectives to decrease the probabilities of certain tokens (e.g., unlikelihood training). However, they often lead to solutions that lack coherence. In this work, we show that an underlying reason for model degeneration is the anisotropic distribution of token representations. We present a contrastive solution: (i) SimCTG, a contrastive training objective to calibrate the model's representation space, and (ii) a decoding method---contrastive search---to encourage diversity while maintaining coherence in the generated text. Extensive experiments and analyses on three benchmarks from two languages demonstrate that our proposed approach outperforms state-of-the-art text generation methods as evaluated by both human and automatic metrics.

## [Polyhisto: Parameter-Efficient Multi-Task Adaptation for Dense Vision Tasks](#)

- Yen-Cheng Liu · CHIH-YAO MA · Junjiao Tian · Zijian He · Zsolt Kira
- abstract@[open-review](#): Adapting large-scale pretrained models to various downstream tasks via fine-tuning is a standard method in machine learning. Recently, parameter-efficient fine-tuning methods have shown promise in adapting a pretrained model to different tasks while training only a few parameters. Despite their success, most existing methods are proposed in Natural Language Processing tasks with language Transformers, and adaptation to Computer Vision tasks with Vision Transformers remains under-explored, especially for dense vision tasks. Further, in multi-task settings, individually fine-tuning and storing separate models for different tasks is inefficient. In this work, we provide an extensive single- and multi-task parameter-efficient benchmark and examine existing parameter-efficient fine-tuning NLP methods for vision tasks. Our results on four different dense vision tasks showed that existing methods cannot be efficiently integrated due to the hierarchical nature of the Hierarchical Vision Transformers. To overcome this issue, we propose Polyhisto and Polyhisto-Lite, consisting of Decomposed HyperNetworks and Layer-wise Scaling Kernels, to share information across different tasks with a few trainable parameters. This leads to favorable performance improvements against existing parameter-efficient methods while using fewer trainable parameters. Specifically, Polyhisto achieves competitive accuracy compared to the state-of-the-art while only using less than 10% of their trainable parameters. Furthermore, our methods show larger performance gains when large networks and more pretraining data are used.

## [Non-deep Networks](#)

- Ankit Goyal · Alexey Bochkovskiy · Jia Deng · Vladlen Koltun
- abstract@[open-review](#): Latency is of utmost importance in safety-critical systems. In neural networks, latency is fundamentally dependent on the depth of the network. This begs the question -- is it possible to build high-performing ``non-deep'' neural networks? We show that it is. To do so, we use parallel subnetworks instead of stacking one layer after another. This helps effectively reduce depth while maintaining high performance. By utilizing parallel substructures, we show, for the first time, that a network with a depth of just 12 can achieve top-1 accuracy over 80% on ImageNet, 96% on CIFAR10, and 81% on CIFAR100. We also show that a network with a low-depth (12) backbone can achieve an AP of 48% on MS-COCO. We analyze the scaling rules for our design and show how to increase performance without changing the network's depth. Finally, we provide a proof of concept for how non-deep networks could be used to build low-latency recognition systems. We will open-source our code.

## [Approximate Secular Equations for the Cubic Regularization Subproblem](#)

- Yihang Gao · Man-Chung Yue · Michael Ng
- abstract@[open-review](#): The cubic regularization method (CR) is a popular algorithm for unconstrained non-convex optimization. At each iteration, CR solves a cubically regularized quadratic problem, called the cubic regularization subproblem (CRS). One way to solve the CRS relies on solving the secular equation, whose computational bottleneck lies in the computation of all eigenvalues of the Hessian matrix. In this paper, we propose and analyze a novel CRS solver based on an approximate secular equation, which requires only some of the Hessian eigenvalues and is therefore much more efficient. Two approximate secular equations (ASEs) are developed. For both ASEs, we first study the existence and uniqueness of their roots and then establish an upper bound on the gap between the root and that of the standard secular equation. Such an upper bound can in turn be used to bound the distance from the approximate CRS solution based ASEs to the true CRS solution, thus offering a theoretical guarantee for our CRS solver. A desirable feature of our CRS solver is that it requires only matrix-vector multiplication but not matrix inversion, which makes it particularly suitable for high-dimensional applications of unconstrained non-convex optimization, such as low-rank recovery and deep learning. Numerical experiments with synthetic and real datasets are conducted to investigate the practical performance of the proposed CRS solver. Experiment results show that the proposed solver outperforms two state-of-the-art methods.

## [Private Estimation with Public Data](#)

- Alex Bie · Gautam Kamath · Vikrant Singhal
- abstract@[open-review](#): We initiate the study of differentially private (DP) estimation with access to a small amount of public data. For private estimation of  $d$ -dimensional Gaussians, we assume that the public data comes from a Gaussian that may have vanishing similarity in TV distance with the underlying Gaussian of the private data. We show that under the constraints of pure or concentrated DP,  $d+1$  public data samples are sufficient to remove any dependence on the range parameters of the private data distribution from the private sample complexity, which is known to be otherwise necessary without public data. For Gaussian mixtures, we assume that the underlying public and private distributions are the same, and we consider two settings: (1) when given a dimension-independent amount of public data, the private sample complexity can be improved polynomially in terms of the number of mixture components, and any dependence on the range parameters of the distribution can be removed in the approximate DP case; (2) when given an amount of public data linear in the dimension, the private sample complexity can be made independent of range parameters even under concentrated DP, and additional improvements can be made to the overall sample complexity.

## [Robust Testing in High-Dimensional Sparse Models](#)

- Anand Jerry George · Clément L Canonne
- abstract@[open-review](#): We consider the problem of robustly testing the norm of a high-dimensional sparse signal vector under two different observation models. In the first model, we are given  $n$  i.i.d. samples from the distribution  $\mathcal{N}(\theta, I_d)$  (with unknown  $\theta$ ), of which a small fraction has been arbitrarily corrupted. Under the promise that  $\|\theta\|_0 \leq s$ , we want to correctly distinguish whether  $\|\theta\|_2 = 0$  or  $\|\theta\|_2 > \gamma$ , for some input parameter  $\gamma > 0$ . We show that any algorithm for this task requires  $n = \Omega(\log(\frac{ed}{s}))$  samples, which is tight up to logarithmic factors. We also extend our results to other common notions of sparsity, namely,  $\|\theta\|_q \leq s$  for any  $0 < q < 2$ . In the second observation model that we consider, the data is generated according to a sparse linear regression model, where the covariates are i.i.d. Gaussian and the regression coefficient (signal) is known to be  $s$ -sparse. Here too we assume that an  $\epsilon$ -fraction of the data is arbitrarily corrupted. We show that any algorithm that reliably tests the norm of the regression coefficient requires at least  $n = \Omega(\min(s \log d, \frac{1}{\gamma^4}))$  samples. Our results show that the complexity of testing in these two settings significantly increases under robustness constraints. This is in line with the recent observations made in robust mean testing and robust covariance testing.

## [Learning Physical Dynamics with Subequivariant Graph Neural Networks](#)

- Jiaqi Han · Wenbing Huang · Hengbo Ma · Jiachen Li · Josh Tenenbaum · Chuang Gan

- abstract@[open-review](#): Graph Neural Networks (GNNs) have become a prevailing tool for learning physical dynamics. However, they still encounter several challenges: 1) Physical laws abide by symmetry, which is a vital inductive bias accounting for model generalization and should be incorporated into the model design. Existing simulators either consider insufficient symmetry, or enforce excessive equivariance in practice when symmetry is partially broken by gravity. 2) Objects in the physical world possess diverse shapes, sizes, and properties, which should be appropriately processed by the model. To tackle these difficulties, we propose a novel backbone, called Subequivariant Graph Neural Network, which 1) relaxes equivariance to subequivariance by considering external fields like gravity, where the universal approximation ability holds theoretically; 2) introduces a new subequivariant object-aware message passing for learning physical interactions between multiple objects of various shapes in particle-based representation; 3) operates in a hierarchical fashion, allowing for modeling long-range and complex interactions. Our model achieves on average over 3% enhancement in contact prediction accuracy across 8 scenarios on Physion and 2\$ lower rollout MSE on RigidFall compared with state-of-the-art GNN simulators, while exhibiting strong generalization and data efficiency.

## [Rethinking training of 3D GANs](#)

- Ivan Skorokhodov · Sergey Tulyakov · Yiqun Wang · Peter Wonka
- abstract@[open-review](#): We are witnessing a surge of works on building and improving 3D-aware generators. To induce a 3D-aware bias, such models rely on volumetric rendering, which is expensive to employ at high resolutions. The dominant strategy to address the scaling issue is to train a separate 2D decoder to upsample a low-resolution volumetrically rendered representation. But this solution comes at a cost. Not only does it break multi-view consistency, e.g. shape and texture change when a camera moves, but it also learns the geometry in a low fidelity. In this work, we take a different route to 3D synthesis and develop a non-upampler-based generator with state-of-the-art image quality, high-resolution geometry and which trains \$2.5 \times\$ faster. For this, we revisit and improve patch-based optimization in two ways. First, we design a location- and scale-aware discriminator by modulating its filters with a hypernetwork. Second, we modify the patch sampling strategy based on an annealed beta distribution to stabilize training and accelerate the convergence. We train on four datasets (two introduced in this work) at \$256^2\$ and \$512^2\$ resolutions, directly, without the need of a 2D upsampler, and our model attains better or comparable FID and has higher fidelity geometry than the current SotA. Code/data/visualizations: <https://rethinking-3d-gans.github.io>

## [Near-Optimal Collaborative Learning in Bandits](#)

- Clémence Rabaud · Sattar Vakili · Emilie Kaufmann
- abstract@[open-review](#): This paper introduces a general multi-agent bandit model in which each agent is facing a finite set of arms and may communicate with other agents through a central controller in order to identify -in pure exploration- or play -in regret minimization- its optimal arm. The twist is that the optimal arm for each agent is the arm with largest expected mixed reward, where the mixed reward of an arm is a weighted sum of the rewards of this arm for all agents. This makes communication between agents often necessary. This general setting allows to recover and extend several recent models for collaborative bandit learning, including the recently proposed federated learning with personalization [Shi et al., 2021]. In this paper, we provide new lower bounds on the sample complexity of pure exploration and on the regret. We then propose a near-optimal algorithm for pure exploration. This algorithm is based on phased elimination with two novel ingredients: a data-dependent sampling scheme within each phase, aimed at matching a relaxation of the lower bound.

## [Decentralized Local Stochastic Extra-Gradient for Variational Inequalities](#)

- Aleksandr Beznosikov · Pavel Dvurechenskii · Anastasiia Koloskova · Valentin Samokhin · Sebastian Stich · Alexander Gasnikov
- abstract@[open-review](#): We consider distributed stochastic variational inequalities (VIs) on unbounded domains with the problem data that is heterogeneous (non-IID) and distributed across many devices. We make a very general assumption on the computational network that, in particular, covers the settings of fully decentralized calculations with time-varying networks and centralized topologies commonly used in Federated Learning. Moreover, multiple local updates on the workers can be made for reducing the communication frequency between workers. We extend the stochastic extragradient method to this very general setting and theoretically analyze its convergence rate in the strongly monotone, monotone, and non-monotone settings when a Minty solution exists. The provided rates explicitly exhibit the dependence on network characteristics (e.g., mixing time), iteration counter, data heterogeneity, variance, number of devices, and other standard parameters. As a special case, our method and analysis apply to distributed stochastic saddle-point problems (SPP), e.g., to training Deep Generative Adversarial Networks (GANs) for which decentralized training has been reported to be extremely challenging. In experiments for decentralized training of GANs we demonstrate the effectiveness of our proposed approach.

## [Multi-Instance Causal Representation Learning for Instance Label Prediction and Out-of-Distribution Generalization](#)

- Weijia Zhang · Xuanhui Zhang · hanwen deng · Min-Ling Zhang
- abstract@[open-review](#): Multi-instance learning (MIL) deals with objects represented as bags of instances and can predict instance labels from bag-level supervision. However, significant performance gaps exist between instance-level MIL algorithms and supervised learners since the instance labels are unavailable in MIL. Most existing MIL algorithms tackle the problem by treating multi-instance bags as harmful ambiguities and predicting instance labels by reducing the supervision inexactness. This work studies MIL from a new perspective by considering bags as auxiliary information, and utilize it to identify instance-level causal representations from bag-level weak supervision. We propose the CausalMIL algorithm, which not only excels at instance label prediction but also provides robustness to distribution change by synergistically integrating MIL with identifiable variational autoencoder. Our approach is based on a practical and general assumption: the prior distribution over the instance latent representations belongs to the non-factorized exponential family conditioning on the multi-instance bags. Experiments on synthetic and real-world datasets demonstrate that our approach significantly outperforms various baselines on instance label prediction and out-of-distribution generalization tasks.

## [Scalable Infomin Learning](#)

- Yanzhi Chen · weihao sun · Yingzhen Li · Adrian Weller
- abstract@[open-review](#): The task of infomin learning aims to learn a representation with high utility while being uninformative about a specified target, with the latter achieved by minimising the mutual information between the representation and the target. It has broad applications, ranging from training fair prediction models against protected attributes, to unsupervised learning with disentangled representations. Recent works on infomin learning mainly use adversarial training, which involves training a neural network to estimate mutual information or its proxy and thus is slow and difficult to optimise. Drawing on recent advances in slicing techniques, we propose a new infomin learning approach, which uses a novel proxy metric to mutual information. We further derive an accurate and analytically computable approximation to this proxy metric, thereby removing the need of constructing neural network-based mutual information estimators. Compared with baselines, experiments on independence tests, disentangled representation learning and fairness tasks demonstrates better performance and higher scalability of our approach.

## [mixReg: A Simple Way to Improve Generalization in Regression for Deep Neural Networks](#)

- Huaxiu Yao · Yiping Wang · Linjun Zhang · James Zou · Chelsea Finn
- abstract@[open-review](#): Improving the generalization of deep networks is an important open challenge, particularly in domains without plentiful data. The mixup algorithm improves generalization by linearly interpolating the input features of a pair of examples and their corresponding labels. These interpolated examples augment the original training dataset. It has shown promising results in various classification tasks, but systematic analysis of mixup in regression remains underexplored. Using mixup directly on regression labels could result in arbitrarily wrong labels since the linearity assumption

behind mixup may not hold. In this paper, we propose a simple yet powerful algorithm, mixReg, to improve generalization on regression tasks. In contrast with the vanilla mixup, which uses the same sampling probability for example pairs, mixReg adjusts the sampling probability based on the similarity of labels. Our theoretical analysis further confirms that mixReg with label similarity obtains a smaller mean square error than vanilla mixup and using feature similarity. Another benefit of mixReg is that it can improve out-of-distribution robustness, where the test distribution is different from the training distribution. By selectively interpolating examples with similar labels, it mitigates the effects of domain-associated information and pushes invariant predictors. We evaluate mixReg on eleven datasets, ranging from tabular to video data. Compared to the best prior approach, mixReg achieves 6.56%, 4.76%, 5.14% improvements in in-distribution generalization, task generalization, and out-of-distribution robustness, respectively.

## [Panchromatic and Multispectral Image Fusion via Alternating Reverse Filtering Network](#)

- Keyu Yan Â· Man Zhou Â· Jie Huang Â· Chengjun Xie Â· Feng Zhao Â· Chongyi Li Â· Danfeng Hong
- abstract@[open-review](#): Panchromatic (PAN) and multi-spectral (MS) image fusion, named Pan-sharpening, refers to super-resolve the low-resolution (LR) multi-spectral (MS) images in the spatial domain to generate the expected high-resolution (HR) MS images, conditioning on the corresponding high-resolution PAN images. In this paper, we present a simple yet effective alternating reverse filtering network for pan-sharpening. Inspired by the classical reverse filtering that reverses images to the status before filtering, we formulate pan-sharpening as an alternately iterative reverse filtering process, which fuses LR MS and HR MS in an interpretable manner. Different from existing model-driven methods that require well-designed priors and degradation assumptions, the reverse filtering process avoids the dependency on pre-defined exact priors. To guarantee the stability and convergence of the iterative process via contraction mapping on a metric space, we develop the learnable multi-scale Gaussian kernel module, instead of using specific filters. We demonstrate the theoretical feasibility of such formulations. Extensive experiments on diverse scenes to thoroughly verify the performance of our method, significantly outperforming the state of the arts. The code will be released.

## [Learning Distinct and Representative Modes for Image Captioning](#)

- Qi Chen Â· Chaorui Deng Â· Qi Wu
- abstract@[open-review](#): Over the years, state-of-the-art (SoTA) image captioning methods have achieved promising results on some evaluation metrics (e.g., CIDEr). However, recent findings show that the captions generated by these methods tend to be biased toward the "average" caption that only captures the most general mode (a.k.a, language pattern) in the training corpus, i.e., the so-called mode collapse problem. Affected by it, the generated captions are limited in diversity and usually less informative than natural image descriptions made by humans. In this paper, we seek to avoid this problem by proposing a Discrete Mode Learning (DML) paradigm for image captioning. Our innovative idea is to explore the rich modes in the training caption corpus to learn a set of "mode embeddings", and further use them to control the mode of the generated captions for existing image captioning models. Specifically, the proposed DML optimizes a dual architecture that consists of an image-conditioned discrete variational autoencoder (CdVAE) branch and a mode-conditioned image captioning (MIC) branch. The CdVAE branch maps each image caption to one of the mode embeddings stored in a learned codebook, and is trained with a pure non-autoregressive generation objective to make the modes distinct and representative. The MIC branch can be simply modified from an existing image captioning model, where the mode embedding is added to the original word embeddings as the control signal. In the experiments, we apply the proposed DML to two widely used image captioning models, Transformer and AoANet. The results show that the learned mode embedding successfully facilitates these models to generate high-quality image captions with different modes, further leading to better performance for both diversity and quality on the MS COCO dataset.

## [LASSIE: Learning Articulated Shapes from Sparse Image Ensemble via 3D Part Discovery](#)

- Chun-Han Yao Â· Wei-Chih Hung Â· Yuanzhen Li Â· Michael Rubinstein Â· Ming-Hsuan Yang Â· Varun Jampani
- abstract@[open-review](#): Creating high-quality articulated 3D models of animals is challenging either via manual creation or using 3D scanning tools. Therefore, techniques to reconstruct articulated 3D objects from 2D images are crucial and highly useful. In this work, we propose a practical problem setting of estimating 3D shape and pose of animals given only a few (about 30) in-the-wild images of a particular animal species (say, horse). Contrary to existing works that rely on pre-defined template shapes, we do not assume any form of 2D or 3D ground-truth annotations, nor do we assume any multi-view or temporal information. Our input image ensemble can have animal instances with varying poses, backgrounds, illuminations and also textures. Our key insight is that 3D parts have much more simplistic shapes compared to the overall animal and that the part shapes are robust w.r.t. animal pose articulations. Using these insights, We propose LASSIE, a novel optimization framework that discovers 3D parts in a self-supervised manner using minimal user intervention. A key driving force behind LASSIE is the enforcing of 2D-3D part consistency using self-supervisory deep features. Experiments on Pascal Part and self-collected in-the-wild animal datasets demonstrate considerably better 3D reconstructions as well as both 2D and 3D part discovery compared to prior art.

## [Riemannian Neural SDE: Learning Stochastic Representations on Manifolds](#)

- Sung Woo Park Â· Hyomin Kim Â· Kyungjae Lee Â· Junseok Kwon
- abstract@[open-review](#): In recent years, the neural stochastic differential equation (NSDE) has gained attention for modeling stochastic representations with great success in various types of applications. However, it typically loses expressivity when the data representation is manifold-valued. To address this issue, we suggest a principled method for expressing the stochastic representation with the Riemannian neural SDE (RNSDE), which extends the conventional Euclidean NSDE. Empirical results for various tasks demonstrate that the proposed method significantly outperforms baseline methods.

## [When Does Group Invariant Learning Survive Spurious Correlations?](#)

- Yimeng Chen Â· Ruibin Xiong Â· Zhi-Ming Ma Â· Yanyan Lan
- abstract@[open-review](#): By inferring latent groups in the training data, recent works introduce invariant learning to the case where environment annotations are unavailable. Typically, learning group invariance under a majority/minority split is empirically shown to be effective in improving out-of-distribution generalization on many datasets. However, theoretical guarantee for these methods on learning invariant mechanisms is lacking. In this paper, we reveal the insufficiency of existing group invariant learning methods in preventing classifiers from depending on spurious correlations in the training set. Specifically, we propose two criteria on judging such sufficiency. Theoretically and empirically, we show that existing methods can violate both criteria and thus fail in generalizing to spurious correlation shifts. Motivated by this, we design a new group invariant learning method, which constructs groups with statistical independence tests, and reweights samples by group label proportion to meet the criteria. Experiments on both synthetic and real data demonstrate that the new method significantly outperforms existing group invariant learning methods in generalizing to spurious correlation shifts.

## [Last-Iterate Convergence of Optimistic Gradient Method for Monotone Variational Inequalities](#)

- Eduard Gorbunov Â· Adrien Taylor Â· Gauthier Gidel
- abstract@[open-review](#): The Past Extrageradient (PEG) [Popov, 1980] method, also known as the Optimistic Gradient method, has known a recent gain in interest in the optimization community with the emergence of variational inequality formulations for machine learning. Recently, in the unconstrained case, Golowich et al. [2020] proved that a  $\$O(1/N)$  last-iterate convergence rate in terms of the squared norm of the operator can be achieved for Lipschitz and monotone operators with a Lipchitz Jacobian. In this work, by introducing a novel analysis through potential functions, we show that (i) this  $\$O(1/N)$  last-iterate convergence can be achieved without any assumption on the Jacobian of the operator, and (ii) it can be extended to the constrained case, which was not derived before even under Lipschitzness of the Jacobian. The proof is significantly different from the one known from Golowich et al.

[2020], and its discovery was computer-aided. Those results close the open question of the last iterate convergence of PEG for monotone variational inequalities.

## [A Fast Post-Training Pruning Framework for Transformers](#)

- Woosuk Kwon · Sehoon Kim · Michael Mahoney · Joseph Hassoun · Kurt Keutzer · Amir Gholami
- abstract@[open-review](#): Pruning is an effective way to reduce the huge inference cost of large Transformer models. However, prior work on model pruning requires retraining the model. This can add high cost and complexity to model deployment, making it difficult to use in many practical situations. To address this, we propose a fast post-training pruning framework for Transformers that does not require any retraining. Given a resource constraint and a sample dataset, our framework automatically prunes the Transformer model using structured sparsity methods. To retain high accuracy without retraining, we introduce three novel techniques: (i) a lightweight mask search algorithm that finds which heads and filters to prune based on the Fisher information; (ii) mask rearrangement that complements the search algorithm; and (iii) mask tuning that reconstructs the output activations for each layer. We apply our method to BERT-BASE and DistilBERT, and we evaluate its effectiveness on GLUE and SQuAD benchmarks. Our framework achieves up to 2.0x reduction in FLOPs and 1.56x speedup in inference latency, while maintaining < 1% loss in accuracy. Importantly, our framework prunes Transformers in less than 3 minutes on a single GPU, which is over two orders of magnitude faster than existing pruning approaches that retrain. Our code will be publicly available at GitHub.

## [Locating and Editing Factual Associations in GPT](#)

- Kevin Meng · David Bau · Alex Andonian · Yonatan Belinkov
- abstract@[open-review](#): We analyze the storage and recall of factual associations in autoregressive transformer language models, finding evidence that these associations correspond to localized, directly-editable computations. We first develop a causal intervention for identifying neuron activations that are decisive in a model's factual predictions. This reveals a distinct set of steps in middle-layer feed-forward modules that mediate factual predictions while processing subject tokens. To test our hypothesis that these computations correspond to factual association recall, we modify feed-forward weights to update specific factual associations using Rank-One Model Editing (ROME). We find that ROME is effective on a standard zero-shot relation extraction (zsRE) model-editing task, comparable to existing methods. To perform a more sensitive evaluation, we also evaluate ROME on a new dataset of counterfactual assertions, on which it simultaneously maintains both specificity and generalization, whereas other methods sacrifice one or another. Our results confirm an important role for mid-layer feed-forward modules in storing factual associations and suggest that direct manipulation of computational mechanisms may be a feasible approach for model editing. The code, dataset, visualizations, and an interactive demo notebook are available in the supplemental materials.

## [A Mixture Of Surprises for Unsupervised Reinforcement Learning](#)

- Andrew Zhao · Matthieu Lin · Yangguang Li · Yong-jin Liu · Gao Huang
- abstract@[open-review](#): Unsupervised reinforcement learning aims at learning a generalist policy in a reward-free manner for fast adaptation to downstream tasks. So far, existing methods propose to provide an intrinsic reward based on surprise. Maximizing or minimizing surprise drives the agent to either explore or gain control over its environment. However, both strategies rely on a strong assumption: the entropy of the environment's dynamics is either high or low. This assumption may not always hold in real-world scenarios, where the entropy of the environment's dynamics may be unknown. Hence, choosing between the two objectives is a dilemma. We propose a novel yet simple mixture of policies to address this concern, allowing us to optimize an objective that simultaneously maximizes and minimizes the surprise. Concretely, we train one mixture component whose objective is to maximize the surprise and another whose objective is to minimize the surprise. Hence, our method does not make assumptions about the entropy of the environment's dynamics. We call our method a Mixture Of SurpriseS (MOSS) for unsupervised reinforcement learning. Experimental results show that, surprisingly, our simple method achieves state-of-the-art performance on the URLB benchmark, outperforming previous pure surprise maximization-based objectives. Our code will be made publicly available.

## [Spatial Pruned Sparse Convolution for 3D Object Detection](#)

- Jianhui Liu · Yukang Chen · Xiaoqing Ye · Zhuotao Tian · Xiao Tan · Xiaojuan Qi
- abstract@[open-review](#): 3D scenes are dominated by a large number of background points, which is redundant for the detection task that mainly needs to focus on foreground objects. In this paper, we analyze major components of existing sparse 3D CNNs and find that 3D CNNs ignores the redundancy of data and further amplifies it in the down-sampling process, which brings a huge amount of extra and unnecessary computational overhead. Inspired by this, we propose a new convolution operator named spatial pruned sparse convolution (SPS-Conv), which includes two variants, spatial pruned submanifold sparse convolution (SPSS-Conv) and spatial pruned regular sparse convolution (SPRS-Conv), both of which are based on the idea of dynamically determine crucial areas for performing computations to reduce redundancy. We empirically find that magnitude of features can serve as an important cues to determine crucial areas which get rid of the heavy computations of learning-based methods. The proposed modules can easily be incorporated into existing sparse 3D CNNs without extra architectural modifications. Extensive experiments on the KITTI and nuScenes datasets demonstrate that our method can achieve more than 50% reduction in GFLOPs without compromising the performance.

## [AnimeSR: Learning Real-World Super-Resolution Models for Animation Videos](#)

- Yanze Wu · Xintao Wang · GEN LI · Ying Shan
- abstract@[open-review](#): This paper studies the problem of real-world video super-resolution (VSR) for animation videos, and reveals three key improvements for practical animation VSR. First, recent real-world super-resolution approaches typically rely on degradation simulation using basic operators without any learning capability, such as blur, noise, and compression. In this work, we propose to learn such basic operators from real low-quality animation videos, and incorporate the learned ones into the degradation generation pipeline. Such neural-network-based basic operators could help to better capture the distribution of real degradations. Second, a large-scale high-quality animation video dataset, AVC, is built to facilitate comprehensive training and evaluations for animation VSR. Third, we further investigate an efficient multi-scale network structure. It takes advantage of the efficiency of unidirectional recurrent networks and the effectiveness of sliding-window-based methods. Thanks to the above delicate designs, our method, AnimeSR, is capable of restoring real-world low-quality animation videos effectively and efficiently, achieving superior performance to previous state-of-the-art methods.

## [Robust On-Policy Sampling for Data-Efficient Policy Evaluation](#)

- Ruijie Zhong · Duohan Zhang · Lukas Schäfer · Stefano Albrecht · Josiah Hanna
- abstract@[open-review](#): Reinforcement learning (RL) algorithms are often categorized as either on-policy or off-policy depending on whether they use data from a target policy of interest or from a different behavior policy. In this paper, we study a subtle distinction between on-policy data and on-policy sampling in the context of the RL sub-problem of policy evaluation. We observe that on-policy sampling may fail to match the expected distribution of on-policy data after observing only a finite number of trajectories and this failure hinders data-efficient policy evaluation. Towards improved data-efficiency, we show how non-i.i.d., off-policy sampling can produce data that more closely matches the expected on-policy data distribution and consequently increases the accuracy of the Monte Carlo estimator for policy evaluation. We introduce a method called Robust On-policy Sampling and demonstrate theoretically and empirically that it produces data that converges faster to the expected on-policy distribution compared to on-policy sampling. Empirically, we show that this faster convergence leads to lower mean squared error policy value estimates.

## [OTKGE: Multi-modal Knowledge Graph Embeddings via Optimal Transport](#)

- Zongsheng Cao · Qianqian Xu · Zhiyong Yang · Yuan He · Xiaochun Cao · Qingming Huang
- abstract@[open-review](#): Multi-modal knowledge graph embeddings (KGE) have shown great power in learning representations of entities and relations for downstream tasks. Different from previous uni-modal KGE approaches, multi-modal KGE can leverage a wealth of multi-modal (textual, visual) knowledge and learn more realistic representations of real-world entities. However, the critical challenge along this course lies in that the multi-modal embedding spaces are usually heterogeneous, and direct fusion will destroy the inherent spatial structure of different modal embeddings, which may harm the interaction of multi-modal knowledge. To overcome this challenge, we innovatively revisit multi-modal KGE from a geometric perspective and propose optimal transport knowledge graph embeddings (OTKGE). Specifically, we model the multi-modal fusion procedure as a transport plan moving different modal embeddings to a unified aligned space by minimizing the Wasserstein distance between multi-modal distributions. Theoretically, we show the distribution differences between source multi-modal spaces and the unified space can be bounded by the Wasserstein distance and demonstrate the advantage of multi-modal KGE in generalization performance over uni-modal KGE. Experimental results on both well-established uni-modal and multi-modal knowledge graph completion benchmarks show that our OTKGE achieves the state-of-the-art performance.

## [Neural Reflectance Field from Shading and Shadow under a Fixed Viewpoint](#)

- Wenqi Yang · Guanying Chen · Chaofeng Chen · Zhenfang Chen · Kwan-Yee K. Wong
- abstract@[open-review](#): In this paper, we address the "dual problem" of multi-view scene reconstruction in which we utilize single-view images captured under different point lights to learn a neural scene representation. Different from existing single-view methods which can only recover a 2.5D scene representation (i.e., a normal / depth map for the visible surface), our method learns a neural reflectance field to represent the 3D geometry and BRDFs of a scene. Instead of relying on multi-view photo-consistency, our method exploits two information-rich monocular cues, namely shading and shadow, to infer scene geometry. Experiments on multiple challenging datasets show that our method is capable of recovering 3D geometry, including both visible and invisible parts, of a scene from single-view images. Thanks to the neural reflectance field representation, our method is robust to depth discontinuities. It supports applications like novel-view synthesis and relighting. Our code and model will be made publicly available.

## [Neur2SP: Neural Two-Stage Stochastic Programming](#)

- Rahul Mihir Patel · Justin Dumouchelle · Elias Khalil · Merve Bodur
- abstract@[open-review](#): Stochastic programming is a powerful modeling framework for decision-making under uncertainty. In this work, we tackle two-stage stochastic programs (2SPs), the most widely applied and studied stochastic programming models. Solving 2SP can take prohibitively long time, especially when the second-stage problem is a mixed-integer linear program (MIP) or a nonlinear program (NLP), even if specialized algorithms that exploit problem structure are employed. Finding high-quality (first-stage) solutions quickly can be crucial in such settings. For this aim, we develop Neur2SP, a new method that approximates the expected (second-stage) value function via a neural network to obtain a surrogate model, which can be solved more efficiently than the original 2SP. The proposed approach makes no assumptions about the problem structure, in particular about the second-stage problem, and can be implemented using an off-the-shelf solver and open-source libraries. Our extensive computational experiments on the benchmark instances of a variety of problem classes, 2SPs with different structures, show the efficiency and efficacy of Neur2SP.

## [Prune and distill: similar reformatting of image information along rat visual cortex and deep neural networks](#)

- Paolo Muratore · Sina Tafazoli · Eugenio Piasini · Alessandro Laio · Davide Zoccolan
- abstract@[open-review](#): Visual object recognition has been extensively studied in both neuroscience and computer vision. Recently, the most popular class of artificial systems for this task, deep convolutional neural networks (CNNs), has been shown to provide excellent models for its functional analogue in the brain, the ventral stream in visual cortex. This has prompted questions on what, if any, are the common principles underlying the reformatting of visual information as it flows through a CNN or the ventral stream. Here we consider some prominent statistical patterns that are known to exist in the internal representations of either CNNs or the visual cortex and look for them in the other system. We show that intrinsic dimensionality (ID) of object representations along the rat homologue of the ventral stream presents two distinct expansion-contraction phases, as previously shown for CNNs. Conversely, in CNNs, we show that training results in both distillation and active pruning (mirroring the increase in ID) of low- to middle-level image information in single units, as representations gain the ability to support invariant discrimination, in agreement with previous observations in rat visual cortex. Taken together, our findings suggest that CNNs and visual cortex share a similarly tight relationship between dimensionality expansion/reduction of object representations and reformatting of image information.

## [Palm up: Playing in the Latent Manifold for Unsupervised Pretraining](#)

- Hao Liu · Tom Zahavy · Volodymyr Mnih · Satinder Singh
- abstract@[open-review](#): Large and diverse datasets have been the cornerstones of many impressive advancements in artificial intelligence. Intelligent creatures, however, learn by interacting with the environment, which changes the input sensory signals and the state of the environment. In this work, we aim to bring the best of both worlds and propose an algorithm that exhibits an exploratory behavior whilst it utilizes large diverse datasets. Our key idea is to leverage deep generative models that are pretrained on static datasets and introduce a dynamic model in the latent space. The transition dynamics simply mixes an action and a random sampled latent. It then applies an exponential moving average for temporal persistency, the resulting latent is decoded to image using pretrained generator. We then employ an unsupervised reinforcement learning algorithm to explore in this environment and perform unsupervised representation learning on the collected data. We further leverage the temporal information of this data to pair data points as a natural supervision for representation learning. Our experiments suggest that the learned representations can be successfully transferred to downstream tasks in both vision and reinforcement learning domains.

## [A Unified Analysis of Mixed Sample Data Augmentation: A Loss Function Perspective](#)

- Chanwoo Park · Sangdoo Yun · Sanghyuk Chun
- abstract@[open-review](#): We propose the first unified theoretical analysis of mixed sample data augmentation (MSDA), such as Mixup and CutMix. Our theoretical results show that regardless of the choice of the mixing strategy, MSDA behaves as a pixel-level regularization of the underlying training loss and a regularization of the first layer parameters. Similarly, our theoretical results support that the MSDA training strategy can improve adversarial robustness and generalization compared to the vanilla training strategy. Using the theoretical results, we provide a high-level understanding of how different design choices of MSDA work differently. For example, we show that the most popular MSDA methods, Mixup and CutMix, behave differently, e.g., CutMix regularizes the input gradients by pixel distances, while Mixup regularizes the input gradients regardless of pixel distances. Our theoretical results also show that the optimal MSDA strategy depends on tasks, datasets, or model parameters. From these observations, we propose generalized MSDAs, a Hybrid version of Mixup and CutMix (HMix) and Gaussian Mixup (GMix), simple extensions of Mixup and CutMix. Our implementation can leverage the advantages of Mixup and CutMix, while our implementation is very efficient, and the computation cost is almost neglectable as Mixup and CutMix. Our empirical study shows that our HMix and GMix outperform the previous state-of-the-art MSDA methods in CIFAR-100 and ImageNet classification tasks.

## [FairVFL: A Fair Vertical Federated Learning Framework with Contrastive Adversarial Learning](#)

- Tao Qi · Fangzhao Wu · Chuhan Wu · Lingjuan Lyu · Tong Xu · Hao Liao · Zhongliang Yang · Yongfeng Huang · Xing Xie
- abstract@[open-review](#): Vertical federated learning (VFL) is a privacy-preserving machine learning paradigm that can learn models from features distributed on different platforms in a privacy-preserving way. Since in real-world applications the data may contain bias on fairness-sensitive features (e.g., gender), VFL models may inherit bias from training data and become unfair for some user groups. However, existing fair machine learning methods usually rely on the centralized storage of fairness-sensitive features to achieve model fairness, which are usually inapplicable in federated scenarios. In this paper, we propose a fair vertical federated learning framework (FairVFL), which can improve the fairness of VFL models. The core idea of FairVFL is to learn unified and fair representations of samples based on the decentralized feature fields in a privacy-preserving way. Specifically, each platform with fairness-insensitive features first learns local data representations from local features. Then, these local representations are uploaded to a server and aggregated into a unified representation for the target task. In order to learn a fair unified representation, we send it to each platform storing fairness-sensitive features and apply adversarial learning to remove bias from the unified representation inherited from the biased data. Moreover, for protecting user privacy, we further propose a contrastive adversarial learning method to remove private information from the unified representation in server before sending it to the platforms keeping fairness-sensitive features. Experiments on two real-world datasets validate that our method can effectively improve model fairness with user privacy well-protected.

## [One Layer is All You Need](#)

- Yue Bai · Huan Wang · Xu Ma · Yitian Zhang · Zhiqiang Tao · Yun Fu
- abstract@[open-review](#): A deeper network structure generally handles more complicated non-linearity and performs more competitively. Nowadays, advanced network designs often contain a large number of repetitive structures (e.g., Transformer). They empower the network capacity to a new level but also increase the model size inevitably, which is unfriendly to either model restoring or transferring. In this study, we are the first to investigate the representative potential of fixed random weights with limited unique values by iteratively learning different masks, leading to a new paradigm for model compression to diminish the model size. Concretely, we utilize one random initialized layer, accompanied with different masks, to convey different feature mappings and represent repetitive modules in a deep network. As a result, the model can be expressed as  $\text{texit\{one-layer\}}$  with a bunch of masks, which significantly reduce the model storage cost. Furthermore, we enhance our strategy by learning masks for a model filled by padding a given random weights sequence. In this way, our method can further lower the space complexity, especially for models without many repetitive architectures. We validate the potential of leveraging random weights and test our compression paradigm based on different network architectures.

## [Test Time Adaptation via Conjugate Pseudo-labels](#)

- Sachin Goyal · Mingjie Sun · Aditi Raghunathan · J. Zico Kolter
- abstract@[open-review](#): Test-time adaptation (TTA) refers to adapting neural networks to distribution shifts, specifically with just access to unlabeled test samples from the new domain at test-time. Prior TTA methods optimize over unsupervised objectives such as the entropy of model predictions in TENT (Wang et al., 2021), but it is unclear what exactly makes a good TTA loss. In this paper, we start by presenting a surprising phenomenon: if we attempt to \$texit{meta-learn}\$ the best '' possible TTA loss over a wide class of functions, then we recover a function that is \$\\textit{remarkably}\$ similar to (a temperature-scaled version of) the softmax-entropy employed by TENT. This only holds, however, if the classifier we are adapting is trained via cross-entropy loss; if the classifier is trained via squared loss, a different best'' TTA loss emerges. To explain this phenomenon, we analyze test-time adaptation through the lens of the training losses's \$\\textit{convex conjugate}\$. We show that under natural conditions, this (unsupervised) conjugate function can be viewed as a good local approximation to the original supervised loss and indeed, it recovers the ``best'' losses found by meta-learning. This leads to a generic recipe than be used to find a good TTA loss for \$\\textit{any}\$ given supervised training loss function of a general class. Empirically, our approach dominates other TTA alternatives over a wide range of domain adaptation benchmarks. Our approach is particularly of interest when applied to classifiers trained with \$\\textit{novel}\$ loss functions, e.g., the recently-proposed PolyLoss (Leng et al., 2022) function, where it differs substantially from (and outperforms) an entropy-based loss. Further, we show that our conjugate based approach can also be interpreted as a kind of self-training using a very specific soft label, which we refer to as the \$\\textit{conjugate pseudo-label}\$. Overall, therefore, our method provides a broad framework for better understanding and improving test-time adaptation.

## [HyperTree Proof Search for Neural Theorem Proving](#)

- Guillaume Lample · Timothee Lacroix · Marie-Anne Lachaux · Aurelien Rodriguez · Amaury Hayat · Thibaut Lavril · Gabriel Ebner · Xavier Martinet
- abstract@[open-review](#): We propose an online training procedure for a transformer-based automated theorem prover. Our approach leverages a new search algorithm, HyperTree Proof Search (HTPS), inspired by the recent success of AlphaZero. Our model learns from previous proof searches through online training, allowing it to generalize to domains far from the training distribution. We report detailed ablations of our pipelineâ€™s main components by studying performance on three environments of increasing complexity. In particular, we show that with HTPS alone, a model trained on annotated proofs manages to prove 65.4% of a held-out set of Metamath theorems, significantly outperforming the previous state of the art of 56.5% by GPT-f. Online training on these unproved theorems increases accuracy to 82.6%. With a similar computational budget, we improve the state of the art on the Lean-based miniF2F-curriculum dataset from 31% to 42% proving accuracy.

## [Learning Options via Compression](#)

- Yiding Jiang · Evan Liu · Benjamin Eysenbach · J. Zico Kolter · Chelsea Finn
- abstract@[open-review](#): Identifying statistical regularities in solutions to some tasks in multi-task reinforcement learning can accelerate learning new tasks. Skill learning offers one way of extracting these regularities by decomposing pre-collected experience into a sequence of skills. A popular approach to skill learning is maximizing the likelihood of the pre-collected experience with latent variable models. However, there are often many different solutions that maximize the likelihood equally well, including degenerate solutions. To address this underspecification, we propose a new objective that combines the maximum likelihood objective with a penalty on the description length of the skills. This penalty incentivizes the skills to maximally identify and extract common structure from the experiences. We demonstrate the effectiveness of our method on a multi-task benchmark from prior work. We demonstrate the effectiveness of our method on a multi-task benchmark from prior work. Further, while most prior works in the offline multi-task setting focus on low-dimensional tasks, we demonstrate that our method can scale to challenging tasks with image observations. Additionally, the acquired skills can be used to solve downstream tasks with up to 8x fewer samples, as compared with skills acquired through maximizing likelihood.

## [Signal Recovery with Non-Expansive Generative Network Priors](#)

- Jorio Cocola
- abstract@[open-review](#): We study compressive sensing with a deep generative network prior. Initial theoretical guarantees for efficient recovery from compressed linear measurements have been developed for signals in the range of a ReLU network with Gaussian weights and logarithmic expansivity: that is when each layer is larger than the previous one by a logarithmic factor. It was later shown that constant expansivity is sufficient for recovery. It has remained open whether the expansivity can be relaxed, allowing for networks with contractive layers (as often the case of real generators). In this work we answer this question, proving that a signal in the range of a Gaussian generative network can be recovered from few linear measurements provided that the width of the layers is proportional to the input layer size (up to log factors). This condition allows the generative network to have contractive layers. Our result is based on showing that Gaussian matrices satisfy a matrix concentration inequality which we term Range Restricted Weight Distribution Condition (R2WDC) and which weakens the Weight Distribution Condition (WDC) upon which previous theoretical guarantees were based. The WDC has also been

used to analyze other signal recovery problems with generative network priors. By replacing the WDC with the R2WDC, we are able to extend previous results for signal recovery with expansive generative network priors to non-expansive ones. We discuss these extensions for phase retrieval, denoising, and spiked matrix recovery.

## [A Continuous Time Framework for Discrete Denoising Models](#)

- Andrew Campbell · Joe Benton · Valentin De Bortoli · Thomas Rainforth · George Deligiannidis · Arnaud Doucet
- abstract@[open-review](#): We provide the first complete continuous time framework for denoising diffusion models of discrete data. This is achieved by formulating the forward noising process and corresponding reverse time generative process as Continuous Time Markov Chains (CTMCs). The model can be efficiently trained using a continuous time version of the ELBO. We simulate the high dimensional CTMC using techniques developed in chemical physics and exploit our continuous time framework to derive high performance samplers that we show can outperform discrete time methods for discrete data. The continuous time treatment also enables us to derive a novel theoretical result bounding the error between the generated sample distribution and the true data distribution.

## [Controllable 3D Face Synthesis with Conditional Generative Occupancy Fields](#)

- Keqiang Sun · Shangzhe Wu · Zhaoyang Huang · Ning Zhang · Quan Wang · Hongsheng Li
- abstract@[open-review](#): Capitalizing on the recent advances in image generation models, existing controllable face image synthesis methods are able to generate high-fidelity images with some levels of controllability, e.g., controlling the shapes, expressions, textures, and poses of the generated face images. However, these methods focus on 2D image generative models, which are prone to producing inconsistent face images under large expression and pose changes. In this paper, we propose a new NeRF-based conditional 3D face synthesis framework, which enables 3D controllability over the generated face images by imposing explicit 3D conditions from 3D face priors. At its core is a conditional Generative Occupancy Field (cGOF) that effectively enforces the shape of the generated face to commit to a given 3D Morphable Model (3DMM) mesh. To achieve accurate control over fine-grained 3D face shapes of the synthesized image, we additionally incorporate a 3D landmark loss as well as a volume warping loss into our synthesis algorithm. Experiments validate the effectiveness of the proposed method, which is able to generate high-fidelity face images and shows more precise 3D controllability than state-of-the-art 2D-based controllable face synthesis methods.

## [Can Push-forward Generative Models Fit Multimodal Distributions?](#)

- Antoine Salmona · Valentin De Bortoli · Julie Delon · Agnes Desolneux
- abstract@[open-review](#): Many generative models synthesize data by transforming a standard Gaussian random variable using a deterministic neural network. Among these models are the Variational Autoencoders and the Generative Adversarial Networks. In this work, we call them "push-forward" models and study their expressivity. We formally demonstrate that the Lipschitz constant of these generative networks has to be large in order to fit multimodal distributions. More precisely, we show that the total variation distance and the Kullback-Leibler divergence between the generated and the data distribution are bounded from below by a constant depending on the mode separation and the Lipschitz constant. Since constraining the Lipschitz constants of neural networks is a common way to stabilize generative models, there is a provable trade-off between the ability of push-forward models to approximate multimodal distributions and the stability of their training. We validate our findings on one-dimensional and image datasets and empirically show that diffusion models do not suffer of such limitation.

## [Wavelet Score-Based Generative Modeling](#)

- Florentin Guth · Simon Coste · Valentin De Bortoli · Stephane Mallat
- abstract@[open-review](#): Score-based generative models (SGMs) synthesize new data samples from Gaussian white noise by running a time-reversed Stochastic Differential Equation (SDE) whose drift coefficient depends on some probabilistic score. The discretization of such SDEs typically requires a large number of time steps and hence a high computational cost. This is because of ill-conditioning properties of the score that we analyze mathematically. We show that SGMs can be considerably accelerated, by factorizing the data distribution into a product of conditional probabilities of wavelet coefficients across scales. The resulting Wavelet Score-based Generative Model (WSGM) synthesizes wavelet coefficients with the same number of time steps at all scales, and its time complexity therefore grows linearly with the image size. This is proved mathematically over Gaussian distributions, and shown numerically over physical processes at phase transition and natural image datasets.

## [Inverse Game Theory for Stackelberg Games: the Blessing of Bounded Rationality](#)

- Jibang Wu · Weiran Shen · Fei Fang · Haifeng Xu
- abstract@[open-review](#): Optimizing strategic decisions (a.k.a. computing equilibrium) is key to the success of many non-cooperative multi-agent applications. However, in many real-world situations, we may face the exact opposite of this game-theoretic problem --- instead of prescribing equilibrium of a given game, we may directly observe the agents' equilibrium behaviors but want to infer the underlying parameters of an unknown game. This research question, also known as inverse game theory, has been studied in multiple recent works in the context of Stackelberg games. Unfortunately, existing works exhibit quite negative results, showing statistical hardness and computational hardness, assuming follower's perfectly rational behaviors. Our work relaxes the perfect rationality agent assumption to the classic quantal response model, a more realistic behavior model of bounded rationality. Interestingly, we show that the smooth property brought by such bounded rationality model actually leads to provably more efficient learning of the follower utility parameters in general Stackelberg games. Systematic empirical experiments on synthesized games confirm our theoretical results and further suggest its robustness beyond the strict quantal response model.

## [Trajectory of Mini-Batch Momentum: Batch Size Saturation and Convergence in High Dimensions](#)

- Kiwon Lee · Andrew Cheng · Elliot Paquette · Courtney Paquette
- abstract@[open-review](#): We analyze the dynamics of large batch stochastic gradient descent with momentum (SGD+M) on the least squares problem when both the number of samples and dimensions are large. In this setting, we show that the dynamics of SGD+M converge to a deterministic discrete Volterra equation as dimension increases, which we analyze. We identify a stability measurement, the implicit conditioning ratio (ICR), which regulates the ability of SGD+M to accelerate the algorithm. When the batch size exceeds this ICR, SGD+M converges linearly at a rate of  $\mathcal{O}(1/\sqrt{\kappa})$ , matching optimal full-batch momentum (in particular performing as well as a full-batch but with a fraction of the size). For batch sizes smaller than the ICR, in contrast, SGD+M has rates that scale like a multiple of the single batch SGD rate. We give explicit choices for the learning rate and momentum parameter in terms of the Hessian spectra that achieve this performance.

## [Posterior Collapse of a Linear Latent Variable Model](#)

- Zihao Wang · Liu Ziyan
- abstract@[open-review](#): This work identifies the existence and cause of a type of posterior collapse that frequently occurs in the Bayesian deep learning practice. For a general linear latent variable model that includes linear variational autoencoders as a special case, we precisely identify the nature of posterior collapse to be the competition between the likelihood and the regularization of the mean due to the prior. Our result also suggests that posterior collapse may be a general problem of learning for deeper architectures and deepens our understanding of Bayesian deep learning.

## [DeVRF: Fast Deformable Voxel Radiance Fields for Dynamic Scenes](#)

- Jia-Wei Liu · Yan-Pei Cao · Weijia Mao · Wenqiao Zhang · David Junhao Zhang · Jussi Keppo · Ying Shan · Xiaohu Qie · Mike Zheng Shou
- abstract@[open-review](#): Modeling dynamic scenes is important for many applications such as virtual reality and telepresence. Despite achieving unprecedented fidelity for novel view synthesis in dynamic scenes, existing methods based on Neural Radiance Fields (NeRF) suffer from slow convergence (i.e., model training time measured in days). In this paper, we present DeVRF, a novel representation to accelerate learning dynamic radiance fields. The core of DeVRF is to model both the 3D canonical space and 4D deformation field of a dynamic, non-rigid scene with explicit and discrete voxel-based representations. However, it is quite challenging to train such a representation which has a large number of model parameters, often resulting in overfitting issues. To overcome this challenge, we devise a novel static-to-dynamic learning paradigm together with a new data capture setup that is convenient to deploy in practice. This paradigm unlocks efficient learning of deformable radiance fields via utilizing the 3D volumetric canonical space learnt from multi-view static images to ease the learning of 4D voxel deformation field with only few-view dynamic sequences. To further improve the efficiency of our DeVRF and its synthesized novel view's quality, we conduct thorough explorations and identify a set of strategies. We evaluate DeVRF on both synthetic and real-world dynamic scenes with different types of deformation. Experiments demonstrate that DeVRF achieves two orders of magnitude speedup (100x faster) with on-par high-fidelity results compared to the previous state-of-the-art approaches.

## [Multi-modal Grouping Network for Weakly-Supervised Audio-Visual Video Parsing](#)

- Shentong Mo · Yapeng Tian
- abstract@[open-review](#): The audio-visual video parsing task aims to parse a video into modality- and category-aware temporal segments. Previous work mainly focuses on weakly-supervised approaches, which learn from video-level event labels. During training, they do not know which modality perceives and meanwhile which temporal segment contains the video event. Since there is no explicit grouping in the existing frameworks, the modality and temporal uncertainties make these methods suffer from false predictions. For instance, segments in the same category could be predicted in different event classes. Learning compact and discriminative multi-modal subspaces is essential for mitigating the issue. To this end, in this paper, we propose a novel Multi-modal Grouping Network, namely MGN, for explicitly semantic-aware grouping. Specifically, MGN aggregates event-aware unimodal features through unimodal grouping in terms of learnable categorical embedding tokens. Furthermore, it leverages the cross-modal grouping for modality-aware prediction to match the video-level target. Our simple framework achieves improving results against previous baselines on weakly-supervised audio-visual video parsing. In addition, our MGN is much more lightweight, using only 47.2% of the parameters of baselines (17 MB vs. 36 MB).

## [Understanding Benign Overfitting in Gradient-Based Meta Learning](#)

- Lisha Chen · Songtao Lu · Tianyi Chen
- abstract@[open-review](#): Meta learning has demonstrated tremendous success in few-shot learning with limited supervised data. In those settings, the meta model is usually overparameterized. While the conventional statistical learning theory suggests that overparameterized models tend to overfit, empirical evidence reveals that overparameterized meta learning methods still work well - a phenomenon often called ``benign overfitting.'' In an attempt to understand this phenomenon, we focus on the meta learning settings with a challenging bilevel structure that we term the gradient-based meta learning, and analyze its generalization performance under an overparameterized meta linear regression model. While our analysis uses the relatively tractable linear models, our theory contributes to understanding the delicate interplay among data heterogeneity, model adaptation and benign overfitting in gradient-based meta learning tasks. We corroborate our theoretical claims through numerical simulations.

## [Experimental Design for Linear Functionals in Reproducing Kernel Hilbert Spaces](#)

- Mojmir Mutny · Andreas Krause
- abstract@[open-review](#): Optimal experimental design seeks to determine the most informative allocation of experiments to infer an unknown statistical quantity. In this work, we investigate optimal design of experiments for the estimation of linear functionals in reproducing kernel Hilbert spaces (RKHSs). This problem has been extensively studied in the linear regression setting under an estimability condition, which allows estimating parameters without bias. We generalize this framework to RKHSs, and allow for the linear functional to be only approximately inferred, i.e., with a fixed bias. This scenario captures many important modern applications such as estimation of gradient maps, integrals and solutions to differential equations. We provide algorithms for constructing bias-aware designs for linear functionals. We derive non-asymptotic confidence sets for fixed and adaptive designs under sub-Gaussian noise, enabling us to certify estimation with bounded error with high probability.

## [Weakly Supervised Knowledge Distillation for Whole Slide Image Classification](#)

- Linhao Qu · Xiaoyuan Luo · Manning Wang · Zhijian Song
- abstract@[open-review](#): Computer-aided pathology diagnosis based on the classification of Whole Slide Image (WSI) plays an important role in clinical practice, and it is often formulated as a weakly-supervised Multiple Instance Learning (MIL) problem. Existing methods solve this problem from either a bag classification or an instance classification perspective. In this paper, we propose an end-to-end weakly supervised knowledge distillation framework (WENO) for WSI classification, which integrates a bag classifier and an instance classifier in a knowledge distillation framework to mutually improve the performance of both classifiers. Specifically, an attention-based bag classifier is used as the teacher network, which is trained with weak bag labels, and an instance classifier is used as the student network, which is trained using the attention scores obtained from the teacher network as soft pseudo labels for the instances in positive bags. An instance feature extractor is shared between the teacher and the student to further enhance the knowledge exchange between them. In addition, we propose a hard positive instance mining strategy based on the output of the student network to force the teacher network to keep mining hard positive instances. WENO is a plug-and-play framework that can be easily applied to any existing attention-based bag classification methods. Extensive experiments on five datasets demonstrate the efficiency of WENO. Code will be publicly available.

## [GenSDF: Two-Stage Learning of Generalizable Signed Distance Functions](#)

- Gene Chou · Ilya Chugunov · Felix Heide
- abstract@[open-review](#): We investigate the generalization capabilities of neural signed distance functions (SDFs) for learning 3D object representations for unseen and unlabeled point clouds. Existing methods can fit SDFs to a handful of object classes and boast fine detail or fast inference speeds, but do not generalize well to unseen shapes. We introduce a two-stage semi-supervised meta-learning approach that transfers shape priors from labeled to unlabeled data to reconstruct unseen object categories. The first stage uses an episodic training scheme to simulate training on unlabeled data and meta-learns initial shape priors. The second stage then introduces unlabeled data with disjoint classes in a semi-supervised scheme to diversify these priors and achieve generalization. We assess our method on both synthetic data and real collected point clouds. Experimental results and analysis validate that our approach outperforms existing neural SDF methods and is capable of robust zero-shot inference on 100+ unseen classes.

## [Revisiting Graph Contrastive Learning from the Perspective of Graph Spectrum](#)

- Nian Liu · Xiao Wang · Deyu Bo · Chuan Shi · Jian Pei
- abstract@[open-review](#): Graph Contrastive Learning (GCL), learning the node representations by augmenting graphs, has attracted considerable attentions. Despite the proliferation of various graph augmentation strategies, there are still some fundamental questions unclear: what information is essentially learned by GCL? Are there some general augmentation rules behind different augmentations? If so, what are they and what insights can they bring? In this

paper, we answer these questions by establishing the connection between GCL and graph spectrum. By an experimental investigation in spectral domain, we firstly find the General grAph augMEntation (GAME) rule for GCL, i.e., the difference of the high-frequency parts between two augmented graphs should be larger than that of low-frequency parts. This rule reveals the fundamental principle to revisit the current graph augmentations and design new effective graph augmentations. Then we theoretically prove that GCL is able to learn the invariance information by contrastive invariance theorem, together with our GAME rule, for the first time, we uncover that the learned representations by GCL essentially encode the low-frequency information, which explains why GCL works. Guided by this rule, we propose a spectral graph contrastive learning module (SpCo), which is a general and GCL-friendly plug-in. We combine it with different existing GCL models, and extensive experiments well demonstrate that it can further improve the performances of a wide variety of different GCL methods.

## [Product Ranking for Revenue Maximization with Multiple Purchases](#)

- Renzhe Xu · Xingxuan Zhang · Bo Li · Yafeng Zhang · Xiaolong Chen · Peng Cui
- abstract@[open-review](#): Product ranking is the core problem for revenue-maximizing online retailers. To design proper product ranking algorithms, various customer behavior models are proposed to characterize the customers' behaviors when they are provided a list of products. However, existing works assume that each customer purchases at most one product or will keep viewing the product list after purchasing a product, which does not agree with the common practice in real scenarios. We assume that each customer can purchase multiple products at will in this work. To model customers' willingness to view and purchase, we set a random attention span and purchase budget, which determines the maximal amount of products that he/she views and purchase, respectively. Under this setting, we first design an optimal ranking policy when the online retailer can precisely model customers' behaviors. Based on the policy, we further develop UCB-like algorithms with  $\tilde{O}(\sqrt{T})$  regret that estimates customers' behaviors and maximize revenue simultaneously in online settings. Experiments on both synthetic and real-world datasets prove the effectiveness of the proposed algorithms.

## [MsSVT: Mixed-scale Sparse Voxel Transformer for 3D Object Detection on Point Clouds](#)

- Shaocong Dong · lihe Ding · Haiyang Wang · Tingfa Xu · Xinli Xu · Jie Wang · Ziyang Bian · Ying Wang · Jianan Li
- abstract@[open-review](#): 3D object detection from the LiDAR point cloud is fundamental to autonomous driving. Large-scale outdoor scenes usually feature significant variance in instance scales, thus requiring features rich in long-range and fine-grained information to support accurate detection. Recent detectors leverage the power of window-based transformers to model long-range dependencies but tend to blur out fine-grained details. To mitigate this gap, we present a novel Mixed-scale Sparse Voxel Transformer, named MsSVT, which can well capture both types of information simultaneously by the divide-and-conquer philosophy. Specifically, MsSVT explicitly divides attention heads into multiple groups, each in charge of attending to information within a particular range. All groups' output is merged to obtain the final mixed-scale features. Moreover, we provide a novel chessboard sampling strategy to reduce the computational complexity of applying a window-based transformer in 3D voxel space. To improve efficiency, we also implement the voxel sampling and gathering operations sparsely with a hash map. Endowed by the powerful capability and high efficiency of modeling mixed-scale information, our single-stage detector built on top of MsSVT surprisingly outperforms state-of-the-art two-stage detectors on Waymo. Code will soon be available.

## [A2: Efficient Automated Attacker for Boosting Adversarial Training](#)

- Zhuoer Xu · Guanghui Zhu · Changhua Meng · shiwen cui · Zhenzhe Ying · Weiqiang Wang · Ming GU · Yihua Huang
- abstract@[open-review](#): Based on the significant improvement of model robustness by AT (Adversarial Training), various variants have been proposed to further boost the performance. Well-recognized methods have focused on different components of AT (e.g., designing loss functions and leveraging additional unlabeled data). It is generally accepted that stronger perturbations yield more robust models. However, how to generate stronger perturbations efficiently is still missed. In this paper, we propose an efficient automated attacker called A2 to boost AT by generating the optimal perturbations on-the-fly during training. A2 is a parameterized automated attacker to search in the attacker space for the best attacker against the defense model and examples. Extensive experiments across different datasets demonstrate that A2 generates stronger perturbations with low extra cost and reliably improves the robustness of various AT methods against different attacks.

## [Motion Forecasting Transformer with Global Intention Localization and Local Movement Refinement](#)

- Shaoshuai Shi · Li Jiang · Dengxin Dai · Bernt Schiele
- abstract@[open-review](#): Predicting multimodal future behavior of traffic participants is essential for robotic vehicles to make safe decisions. Existing works explore to directly predict future trajectories based on latent features or utilize dense goal candidates to identify agent's destinations, where the former strategy converges slowly since all motion modes are derived from the same feature while the latter strategy has efficiency issue since its performance highly relies on the density of goal candidates. In this paper, we propose the Motion TRansformer (MTR) framework that models motion prediction as the joint optimization of global intention localization and local movement refinement. Instead of using goal candidates, MTR incorporates spatial intention priors by adopting a small set of learnable motion query pairs. Each motion query pair takes charge of trajectory prediction and refinement for a specific motion mode, which stabilizes the training process and facilitates better multimodal predictions. Experiments show that MTR achieves state-of-the-art performance on both the marginal and joint motion prediction challenges, ranking 1st on the leaderboards of Waymo Open Motion Dataset. Code will be available.

## [BYOL-Explore: Exploration by Bootstrapped Prediction](#)

- Zhaohan Guo · Shantanu Thakoor · Miruna Pislar · Bernardo Avila Pires · Florent AltchÃ© · Corentin Tallec · Alaa Saade · Daniele Calandriello · Jean-Bastien Grill · Yunhao Tang · Michal Valko · Remi Munos · Mohammad Gheshlaghi Azar · Bilal Piot
- abstract@[open-review](#): We present BYOL-Explore, a conceptually simple yet general approach for curiosity-driven exploration in visually complex environments. BYOL-Explore learns the world representation, the world dynamics and the exploration policy all-together by optimizing a single prediction loss in the latent space with no additional auxiliary objective. We show that BYOL-Explore is effective in DM-HARD-8, a challenging partially-observable continuous-action hard-exploration benchmark with visually rich 3-D environment. On this benchmark, we solve the majority of the tasks purely through augmenting the extrinsic reward with BYOL-Explore intrinsic reward, whereas prior work could only get off the ground with human demonstrations. As further evidence of the generality of BYOL-Explore, we show that it achieves superhuman performance on the ten hardest exploration games in Atari while having a much simpler design than other competitive agents.

## [Bridging the Gap between Object and Image-level Representations for Open-Vocabulary Detection](#)

- Hanooza Bangalath · Muhammad Maaz · Muhammad Uzair Khattak · Salman Khan · Fahad Shahbaz Khan
- abstract@[open-review](#): Existing open-vocabulary object detectors typically enlarge their vocabulary sizes by leveraging different forms of weak supervision. This helps generalize to novel objects at inference. Two popular forms of weak-supervision used in open-vocabulary detection (OVD) include pretrained CLIP model and image-level supervision. We note that both these modes of supervision are not optimally aligned for the detection task: CLIP is trained with image-text pairs and lacks precise localization of objects while the image-level supervision has been used with heuristics that do not accurately specify local object regions. In this work, we propose to address this problem by performing object-centric alignment of the language embeddings from the CLIP model. Furthermore, we visually ground the objects with only image-level supervision using a pseudo-labeling process that provides high-quality object proposals and helps expand the vocabulary during training. We establish a bridge between the above two object-alignment strategies via a novel weight transfer function that aggregates their complimentary strengths. In essence, the proposed model seeks to minimize the gap

between object and image-centric representations in the OVD setting. On the COCO benchmark, our proposed approach achieves 40.3 AP\$\_{\{50\}}\$ on novel classes, an absolute 11.9 gain over the previous best performance. For LVIS, we surpass the state-of-the-art ViLD model by 5.0 mask AP for rare categories and 3.4 overall. Our codes will be publicly released.

## [Measuring and Reducing Model Update Regression in Structured Prediction for NLP](#)

- Deng Cai · Elman Mansimov · Yi-An Lai · Yixuan Su · Lei Shu · Yi Zhang
- abstract@[open-review](#): Recent advance in deep learning has led to rapid adoption of machine learning based NLP models in a wide range of applications. Despite the continuous gain in accuracy, backward compatibility is also an important aspect for industrial applications, yet it received little research attention. Backward compatibility requires that the new model does not regress on cases that were correctly handled by its predecessor. This work studies model update regression in structured prediction tasks. We choose syntactic dependency parsing and conversational semantic parsing as representative examples of structured prediction tasks in NLP. First, we measure and analyze model update regression in different model update settings. Next, we explore and benchmark existing techniques for reducing model update regression including model ensemble and knowledge distillation. We further propose a simple and effective method, Backward-Congruent Re-ranking (BCR), by taking into account the characteristics of structured output. Experiments show that BCR can better mitigate model update regression than model ensemble and knowledge distillation approaches.

## [CoNSoLe: Convex Neural Symbolic Learning](#)

- Haoran Li · Yang Weng · Hanghang Tong
- abstract@[open-review](#): Learning the underlying equation from data is a fundamental problem in many disciplines. Recent advances rely on Neural Networks (NNs) but do not provide theoretical guarantees in obtaining the exact equations owing to the non-convexity of NNs. In this paper, we propose Convex Neural Symbolic Learning (CoNSoLe) to seek convexity under mild conditions. The main idea is to decompose the recovering process into two steps and convexify each step. In the first step of searching for right symbols, we convexify the deep Q-learning. The key is to maintain double convexity for both the negative Q-function and the negative reward function in each iteration, leading to provable convexity of the negative optimal Q function to learn the true symbol connections. Conditioned on the exact searching result, we construct a Locally Convex equation Learning (LoCaL) neural network to convexify the estimation of symbol coefficients. With such a design, we quantify a large region with strict convexity in the loss surface of LoCaL for commonly used physical functions. Finally, we demonstrate the superior performance of the CoNSoLe framework over the state-of-the-art on a diverse set of datasets.

## [Variance Reduced ProxSkip: Algorithm, Theory and Application to Federated Learning](#)

- Grigory Malinovsky · Kai Yi · Peter Richtarik
- abstract@[open-review](#): We study distributed optimization methods based on the {\em local training (LT)} paradigm, i.e., methods which achieve communication efficiency by performing richer local gradient-based training on the clients before (expensive) parameter averaging is allowed to take place. While these methods were first proposed about a decade ago, and form the algorithmic backbone of federated learning, there is an enormous gap between their practical performance, and our theoretical understanding. Looking back at the progress of the field, we {\em identify} 5 generations of LT methods: 1) heuristic, 2) homogeneous, 3) sublinear, 4) linear, and 5) accelerated. The 5\${\}^{\wedge}\backslash rm th\\$ generation was initiated by the ProxSkip method of Mishchenko et al. (2022), whose analysis provided the first theoretical confirmation that LT is a communication acceleration mechanism. Inspired by this recent progress, we contribute to the 5\${\}^{\wedge}\backslash rm th\\$ generation of LT methods by showing that it is possible to enhance ProxSkip further using {\em variance reduction}. While all previous theoretical results for LT methods ignore the cost of local work altogether, and are framed purely in terms of the number of communication rounds, we construct a method that can be substantially faster in terms of the {\em total training time} than the state-of-the-art method ProxSkip in theory and practice in the regime when local computation is sufficiently expensive. We characterize this threshold theoretically, and confirm our theoretical predictions with empirical results. Our treatment of variance reduction is generic, and can work with a large number of variance reduction techniques, which may lead to future applications in the future. Finally, we corroborate our theoretical results with carefully engineered proof-of-concept experiments.

## [3D Concept Grounding on Neural Fields](#)

- Yining Hong · Yilun Du · Chunru Lin · Josh Tenenbaum · Chuang Gan
- abstract@[open-review](#): In this paper, we address the challenging problem of 3D concept grounding (i.e., segmenting and learning visual concepts) by looking at RGBD images and reasoning about paired questions and answers. Existing visual reasoning approaches typically utilize supervised methods to extract 2D segmentation masks on which concepts are grounded. In contrast, humans are capable of grounding concepts on the underlying 3D representation of images. However, traditionally inferred 3D representations (e.g., point clouds, voxelgrids and meshes) cannot capture continuous 3D features flexibly, thus making it challenging to ground concepts to 3D regions based on the language description of the object being referred to. To address both issues, we propose to leverage the continuous, differentiable nature of neural fields to segment and learn concepts. Specifically, each 3D coordinate in a scene is represented as a high dimensional descriptor. Concept grounding can then be performed by computing the similarity between the descriptor vector of a 3D coordinate and the vector embedding of a language concept, which enables segmentations and concept learning to be jointly learned on neural fields in a differentiable fashion. As a result, both 3D semantic and instance segmentations can emerge directly from question answering supervision using a set of defined neural operators on top of neural fields (e.g., filtering and counting). Experimental results show that our proposed framework outperforms unsupervised / language-mediated segmentation models on semantic and instance segmentation tasks, as well as outperforms existing models on the challenging 3D aware visual reasoning tasks. Furthermore, our framework can generalize well to unseen shape categories and real scans.

## [Improved Fine-Tuning by Better Leveraging Pre-Training Data](#)

- Ziquan Liu · Yi Xu · Yuanhong Xu · Qi Qian · Hao Li · Xiangyang Ji · Antoni Chan · Rong Jin
- abstract@[open-review](#): As a dominant paradigm, fine-tuning a pre-trained model on the target data is widely used in many deep learning applications, especially for small data sets. However, recent studies have empirically shown that training from scratch has the final performance that is no worse than this pre-training strategy once the number of training samples is increased in some vision tasks. In this work, we revisit this phenomenon from the perspective of generalization analysis by using excess risk bound which is popular in learning theory. The result reveals that the excess risk bound may have a weak dependency on the pre-trained model. The observation inspires us to leverage pre-training data for fine-tuning, since this data is also available for fine-tuning. The generalization result of using pre-training data shows that the excess risk bound on a target task can be improved when the appropriate pre-training data is included in fine-tuning. With the theoretical motivation, we propose a novel selection strategy to select a subset from pre-training data to help improve the generalization on the target task. Extensive experimental results for image classification tasks on 8 benchmark data sets verify the effectiveness of the proposed data selection based fine-tuning pipeline.

## [A Coupled Design of Exploiting Record Similarity for Practical Vertical Federated Learning](#)

- Zhaomin Wu · Qinbin Li · Bingsheng He
- abstract@[open-review](#): Federated learning is a learning paradigm to enable collaborative learning across different parties without revealing raw data. Notably, vertical federated learning (VFL), where parties share the same set of samples but only hold partial features, has a wide range of real-world applications. However, most existing studies in VFL disregard the ``record linkage'' process. They design algorithms either assuming the data from

different parties can be exactly linked or simply linking each record with its most similar neighboring record. These approaches may fail to capture the key features from other less similar records. Moreover, such improper linkage cannot be corrected by training since existing approaches provide no feedback on linkage during training. In this paper, we design a novel coupled training paradigm, FedSim, that integrates one-to-many linkage into the training process. Besides enabling VFL in many real-world applications with fuzzy identifiers, FedSim also achieves better performance in traditional VFL tasks. Moreover, we theoretically analyze the additional privacy risk incurred by sharing similarities. Our experiments on eight datasets with various similarity metrics show that FedSim outperforms other state-of-the-art baselines.

## [Clipped Stochastic Methods for Variational Inequalities with Heavy-Tailed Noise](#)

- Eduard Gorbunov · Marina Danilova · David Dobre · Pavel Dvurechenskii · Alexander Gasnikov · Gauthier Gidel
- abstract@[open-review](#): Stochastic first-order methods such as Stochastic Extragradient (SEG) or Stochastic Gradient Descent-Ascent (SGDA) for solving smooth minimax problems and, more generally, variational inequality problems (VIP) have been gaining a lot of attention in recent years due to the growing popularity of adversarial formulations in machine learning. While high-probability convergence bounds are known to more accurately reflect the actual behavior of stochastic methods, most convergence results are provided in expectation. Moreover, the only known high-probability complexity results have been derived under restrictive sub-Gaussian (light-tailed) noise and bounded domain assumptions [Juditsky et al., 2011]. In this work, we prove the first high-probability complexity results with logarithmic dependence on the confidence level for stochastic methods for solving monotone and structured non-monotone VIPs with non-sub-Gaussian (heavy-tailed) noise and unbounded domains. In the monotone case, our results match the best known ones in the light-tails case [Juditsky et al., 2011], and are novel for structured non-monotone problems such as negative comonotone, quasi-strongly monotone, and/or star-cocoercive ones. We achieve these results by studying SEG and SGDA with clipping. In addition, we numerically validate that the gradient noise of many practical GAN formulations is heavy-tailed and show that clipping improves the performance of SEG/SGDA.

## [Phase transitions in when feedback is useful](#)

- Lokesh Boominathan · Xaq Pitkow
- abstract@[open-review](#): Sensory observations about the world are invariably ambiguous. Inference about the world's latent variables is thus an important computation for the brain. However, computational constraints limit the performance of these computations. These constraints include energetic costs for neural activity and noise on every channel. Efficient coding is one prominent theory that describes how such limited resources can best be used. In one incarnation, this leads to a theory of predictive coding, where predictions are subtracted from signals, reducing the cost of sending something that is already known. This theory does not, however, account for the costs or noise associated with those predictions. Here we offer a theory that accounts for both feedforward and feedback costs, and noise in all computations. We formulate this inference problem as message-passing on a graph whereby feedback serves as an internal control signal aiming to maximize how well an inference tracks a target state while minimizing the costs of computation. We apply this novel formulation of inference as control to the canonical problem of inferring the hidden scalar state of a linear dynamical system with Gaussian variability. The best solution depends on architectural constraints, such as Dale's law, the ubiquitous law that each neuron makes solely excitatory or inhibitory postsynaptic connections. This biological structure can create asymmetric costs for feedforward and feedback channels. Under such conditions, our theory predicts the gain of optimal predictive feedback and how it is incorporated into the inference computation. We show that there is a non-monotonic dependence of optimal feedback gain as a function of both the computational parameters and the world dynamics, leading to phase transitions in whether feedback provides any utility in optimal inference under computational constraints.

## [Geodesic Graph Neural Network for Efficient Graph Representation Learning](#)

- Lecheng Kong · Muhan Zhang · Yixin Chen
- abstract@[open-review](#): Recently, Graph Neural Networks (GNNs) have been applied to graph learning tasks and achieved state-of-the-art results. However, many competitive methods employ preprocessing on the target nodes, such as subgraph extraction and customized labeling, to capture some information that is hard to be learned by GNNs. Such operations are time-consuming and do not scale to large graphs. In this paper, we propose an efficient GNN framework called Geodesic GNN (GD-GNN). It injects the conditional relationship between nodes into the model without labeling. Specifically, we view the shortest paths between two nodes as the spatial graph context of the neighborhood around them. The GNN embeddings of nodes on the shortest paths are used to generate geodesic representations. Conditioned on the geodesic representations, GD-GNN is able to generate node, link, and graph representations that carry much richer structural information than plain GNNs. We theoretically prove that GD-GNN is more powerful than plain GNNs, and present experimental results to show that GD-GNN achieves highly competitive performance with state-of-the-art GNN models on link prediction and graph classification tasks while taking significantly less time.

## [On the Strong Correlation Between Model Invariance and Generalization](#)

- Weijian Deng · Stephen Gould · Liang Zheng
- abstract@[open-review](#): Generalization and invariance are two essential properties of machine learning models. Generalization captures a model's ability to classify unseen data while invariance measures consistency of model predictions on transformations of the data. Existing research suggests a positive relationship: a model generalizing well should be invariant to certain visual factors. Building on this qualitative implication we make two contributions. First, we introduce effective invariance (EI), a simple and reasonable measure of model invariance which does not rely on image labels. Given predictions on a test image and its transformed version, EI measures how well the predictions agree and with what level of confidence. Second, using invariance scores computed by EI, we perform large-scale quantitative correlation studies between generalization and invariance, focusing on rotation and grayscale transformations. From a model-centric view, we observe generalization and invariance of different models exhibit a strong linear relationship, on both in-distribution and out-of-distribution datasets. From a dataset-centric view, we find a certain model's accuracy and invariance linearly correlated on different test sets. Apart from these major findings, other minor but interesting insights are also discussed.

## [Unsupervised Multi-Object Segmentation by Predicting Probable Motion Patterns](#)

- Laurynas Karazija · Subhabrata Choudhury · Iro Laina · Christian Rupprecht · Andrea Vedaldi
- abstract@[open-review](#): We propose a new approach to learn to segment multiple image objects without manual supervision. The method can extract objects from still images, but uses videos for supervision. While prior works have considered motion for segmentation, a key insight is that, while motion can be used to identify objects, not all objects are necessarily in motion: the absence of motion does not imply the absence of objects. Hence, our model learns to predict image regions that are likely to contain motion patterns characteristic of objects moving rigidly. It does not predict a specific motion, which cannot be done from a still image, but a distribution of possible motions, which includes the option that an object does not move at all. We demonstrate the advantage of this approach over a deterministic counterpart, show state-of-the-art unsupervised instance segmentation performance on benchmarks, and performance competitive with methods that use motion at test time.

## [Hierarchical Normalization for Robust Monocular Depth Estimation](#)

- Chi Zhang · Wei Yin · Billzb Wang · Gang Yu · Chunhua Shen · BIN FU
- abstract@[open-review](#): In this paper, we address monocular depth estimation with deep neural networks. To enable training of deep monocular estimation models with various sources of datasets, state-of-the-art methods adopt image-level normalization strategies to generate affine-invariant depth representations. However, learning with the image-level normalization mainly emphasizes the relations of pixel representations with the global statistic in the images, such as the structure of the scene, while the fine-grained depth difference may be overlooked. In this paper, we propose a novel multi-scale

depth normalization method that hierarchically normalizes the depth representations based on spatial information and depth distributions. Compared with previous normalization strategies applied only at the holistic image level, the proposed hierarchical normalization can effectively preserve the fine-grained details and improve accuracy. We present two strategies that define the hierarchical normalization contexts in the depth domain and the spatial domain, respectively. Our extensive experiments show that the proposed normalization strategy remarkably outperforms previous normalization methods, and we set new state-of-the-art on five zero-shot transfer benchmark datasets.

## [Estimating Noise Transition Matrix with Label Correlations for Noisy Multi-Label Learning](#)

- Shikun Li · Xiaobo Xia · Hansong Zhang · Yibing Zhan · Shiming Ge · Tongliang Liu
- abstract@[open-review](#): In label-noise learning, the noise transition matrix, bridging the class posterior for noisy and clean data, has been widely exploited to learn statistically consistent classifiers. The effectiveness of these algorithms relies heavily on estimating the transition matrix. Recently, the problem of label-noise learning in multi-label classification has received increasing attention, and these consistent algorithms can be applied in multi-label cases. However, the estimation of transition matrices in noisy multi-label learning has not been studied and remains challenging, since most of the existing estimators in noisy multi-class learning depend on the existence of anchor points and the accurate fitting of noisy class posterior. To address this problem, in this paper, we first study the identifiability problem of the class-dependent transition matrix in noisy multi-label learning, and then inspired by the identifiability results, we propose a new estimator by exploiting label correlations without both anchor points and accurate fitting of noisy class posterior. Specifically, we estimate the occurrence probability of two noisy labels to get noisy label correlations. Then, we perform sample selection to extract side information about clean label correlations, which is used to estimate the occurrence probability of one noisy label when a certain clean label appears. By utilizing the mismatch of label correlations implied in these occurrence probabilities, the transition matrix is identifiable, and can then be inferred by solving a simple bilinear decomposition problem. Empirical results illustrate the effectiveness of our estimator to estimate the transition matrix with label correlations, leading to better classification performance.

## [Decoupling Classifier for Boosting Few-shot Object Detection and Instance Segmentation](#)

- Bin-Bin Gao · Xiaochen Chen · Zhongyi Huang · Congchong Nie · Jun Liu · Jinxiang Lai · GUANNAN JIANG · Xi Wang · Chengjie Wang
- abstract@[open-review](#): This paper focus on few-shot object detection~(FSOD) and instance segmentation~(FSIS), which requires a model to quickly adapt to novel classes with a few labeled instances. The existing methods severely suffer from bias classification because of the missing label issue which naturally exists in a few-shot scenario and is first formally proposed by us. Our analysis suggests that the standard classification head of most FSOD or FSIS models needs to be decoupled to mitigate the bias classification. Therefore, we propose an embarrassingly simple but effective method that decouples the standard classifier into two heads. Then, these two individual heads are capable of independently addressing clear positive samples and noisy negative samples which are caused by the missing label. In this way, the model can effectively learn novel classes while mitigating the effects of noisy negative samples. Without bells and whistles, our model without any additional computation cost and parameters consistently outperforms its baseline and state-of-the-art by a large margin on PASCAL VOC and MS-COCO benchmarks for FSOD and FSIS tasks. The code will be available.

## [Don't Throw Your Model Checkpoints Away](#)

- Chaofei Wang · Qisen Yang · Rui Huang · Shiji Song · Gao Huang
- abstract@[open-review](#): Knowledge distillation is an effective approach to learn compact models (students) with the supervision of large and strong models (teachers). As empirically there exists a strong correlation between the performance of teacher and student models, it is commonly believed that a high performing teacher is preferred. Consequently, practitioners tend to use a well trained network or an ensemble of them as the teacher. In this paper, we make an intriguing observation that an intermediate model, i.e., a checkpoint in the middle of the training procedure, often serves as a better teacher compared to the fully converged model, although the former has much lower accuracy. More surprisingly, a weak snapshot ensemble of several intermediate models from a same training trajectory can outperform a strong ensemble of independently trained and fully converged models, when they are used as teachers. We show that this phenomenon can be partially explained by the information bottleneck principle: the feature representations of intermediate models can have higher mutual information regarding the input, and thus contain more "dark knowledge" for effective distillation. We further propose an optimal intermediate teacher selection algorithm based on maximizing the total task-related mutual information. Experiments verify its effectiveness and applicability.

## [TOIST: Task Oriented Instance Segmentation Transformer with Noun-Pronoun Distillation](#)

- Pengfei Li · Beiwen Tian · Yongliang Shi · Xiaoxue Chen · Hao Zhao · Guyue Zhou · Ya-Qin Zhang
- abstract@[open-review](#): Current referring expression comprehension algorithms can effectively detect or segment objects indicated by nouns, but how to understand verb reference is still under-explored. As such, we study the challenging problem of task oriented detection, which aims to find objects that best afford an action indicated by verbs like sit comfortably on. Towards a finer localization that better serves downstream applications like robot interaction, we extend the problem into task oriented instance segmentation. A unique requirement of this task is to select preferred candidates among possible alternatives. Thus we resort to the transformer architecture which naturally models pair-wise query relationships with attention, leading to the TOIST method. In order to leverage pre-trained noun referring expression comprehension models and the fact that we can access privileged noun ground truth during training, a novel noun-pronoun distillation framework is proposed. Noun prototypes are generated in an unsupervised manner and contextual pronoun features are trained to select prototypes. As such, the network remains noun-agnostic during inference. We evaluate TOIST on the large-scale task oriented dataset COCO-Tasks and achieve +10.7% higher  $\text{mAP}^{\{\text{box}\}}$  than the best-reported results. The proposed noun-pronoun distillation can boost  $\text{mAP}^{\{\text{box}\}}$  and  $\text{mAP}^{\{\text{mask}\}}$  by +2.6% and +3.6%. Codes and models are publicly available.

## [Basis Encoded Polynomial Neural Fields for Subband Decomposition](#)

- Guandao Yang · Sagie Benaim · Varun Jampani · Kyle Genova · Jonathan Barron · Thomas Funkhouser · Serge Belongie · Bharath Hariharan
- abstract@[open-review](#): Neural fields have emerged as a new paradigm for representing signals, thanks to their ability to represent signals compactly while being easy to optimize. In most applications, however, neural fields are treated like a black box, which precludes many signal manipulation tasks. In this paper, we propose a new class of neural fields called basis-encoded polynomial neural fields (PNFs). The key advantage of a PNF is that it can represent a signal as a composition of a number of manipulable and interpretable components without losing the merits of neural fields representation. We develop a general theoretical framework to analyze and design PNFs. We use this framework to design Fourier PNFs, which match state-of-the-art performance in signal representation tasks that use neural fields. In addition, we empirically demonstrate that Fourier PNFs enable signal manipulation applications such as texture transfer and scale-space interpolation.

## [Predicting from Predictions](#)

- Frances Ding · Yixin Wang · Celestine Mendler-Dünner
- abstract@[open-review](#): Predictions about people, such as their expected educational achievement or their credit risk, can shape the outcome that they aim to predict. Estimating the causal effect of these predictions is important for deciding which predictive models to deploy. However, this estimation poses unique challenges because model predictions are usually deterministic functions of model inputs (making the effects of predictions difficult to disentangle from the effects of other inputs), and predictions are also highly correlated with outcomes. In this work we show that any of the three following conditions are sufficient to identify the causal effect of predictions: overparameterization of the predictive model, randomization in released predictions, and measurement noise. Under these conditions, standard supervised learning implicitly estimates the causal effect of predictions and finds transferable

functional relationships to anticipate outcomes from a new model deployment, as long as model predictions are provided as a feature. Empirically, we find that supervised learning succeeds even in the presence of mild violations of our assumptions, such as finite data and model misspecification. Lastly, we discuss how this approach, unlike standard individualistic models, is able to handle certain types of interference (people affecting the outcomes of others). These findings emphasize the importance of recording model predictions during deployment, in order to understand social outcomes and feedback loops.

## [DAGMA: Learning DAGs via M-matrices and a Log-Determinant Acyclicity Characterization](#)

- Kevin Bello · Bryon Aragam · Pradeep Ravikumar
- abstract@[open-review](#): The combinatorial problem of learning directed acyclic graphs (DAGs) from data was recently framed as a purely continuous optimization problem by leveraging a differentiable acyclicity characterization of DAGs based on the trace of a matrix exponential function. Existing acyclicity characterizations are based on the idea that powers of an adjacency matrix contain information about walks and cycles. In this work, we propose a fundamentally different acyclicity characterization based on the log-determinant (log-det) function, which leverages the nilpotency property of DAGs. To deal with the inherent asymmetries of a DAG, we relate the domain of our log-det characterization to the set of M-matrices, which is a key difference to the classical log-det function defined over the cone of positive definite matrices. Similar to acyclicity functions previously proposed, our characterization is also exact and differentiable. However, when compared to existing characterizations, our log-det function: (1) Is better at detecting large cycles; (2) Has better behaved gradients; and (3) Its runtime is in practice about an order of magnitude faster. From the optimization side, we drop the typically used augmented Lagrangian scheme, and propose DAGMA (Direct Acyclic Graphs via M-matrices for Acyclicity), a method that resembles the central path approach for barrier methods. Each point in the central path of DAGMA is a solution to an unconstrained problem regularized by our log-det function, then we show that at the limit of the central path, the solution is guaranteed to be a DAG. Finally, we provide extensive experiments for linear and nonlinear SEMs, and show that our approach can reach large speed-ups and smaller structural Hamming distance against state-of-the-art methods.

## [Exploring the Algorithm-Dependent Generalization of AUPRC Optimization with List Stability](#)

- Peisong Wen · Qianqian Xu · Zhiyong Yang · Yuan He · Qingming Huang
- abstract@[open-review](#): Stochastic optimization of the Area Under the Precision-Recall Curve (AUPRC) is a crucial problem for machine learning. Although various algorithms have been extensively studied for AUPRC optimization, the generalization is only guaranteed in the multi-query case. In this work, we present the first trial in the single-query generalization of stochastic AUPRC optimization. For sharper generalization bounds, we focus on algorithm-dependent generalization. There are both algorithmic and theoretical obstacles to our destination. From an algorithmic perspective, we notice that the majority of existing stochastic estimators are unbiased only when the sampling strategy is unbiased, and is leave-one-out unstable due to the non-decomposability. To address these issues, we propose a sampling-rate-invariant unbiased stochastic estimator with superior stability. On top of this, the AUPRC optimization is formulated as a composition optimization problem, and a stochastic algorithm is proposed to solve this problem. From a theoretical perspective, standard techniques of the algorithm-dependent generalization analysis cannot be directly applied to such a listwise compositional optimization problem. To fill this gap, we extend the model stability from instancewise losses to listwise losses and bridge the corresponding generalization and stability. Additionally, we construct state transition matrices to describe the recurrence of the stability, and simplify calculations by matrix spectrum. Practically, experimental results on three real-world datasets speak to the effectiveness and soundness of our framework.

## [Minimax Optimal Fixed-Budget Best Arm Identification in Linear Bandits](#)

- Junwen Yang · Vincent Tan
- abstract@[open-review](#): We study the problem of best arm identification in linear bandits in the fixed-budget setting. By leveraging properties of the G-optimal design and incorporating it into the arm allocation rule, we design a parameter-free algorithm, Optimal Design-based Linear Best Arm Identification (OD-LinBAI). We provide a theoretical analysis of the failure probability of OD-LinBAI. Instead of all the optimality gaps, the performance of OD-LinBAI depends only on the gaps of the top  $\$d\$$  arms, where  $\$d\$$  is the effective dimension of the linear bandit instance. Complementarily, we present a minimax lower bound for this problem. The upper and lower bounds show that OD-LinBAI is minimax optimal up to constant multiplicative factors in the exponent, which is a significant theoretical improvement over existing methods (e.g., BayesGap, Peace, LinearExploration and GSE), and settles the question of ascertaining the difficulty of learning the best arm in the fixed-budget setting. Finally, numerical experiments demonstrate considerable empirical improvements over existing algorithms on a variety of real and synthetic datasets.

## [LOT: Layer-wise Orthogonal Training on Improving L2 Certified Robustness](#)

- Xiaojun Xu · Linyi Li · Bo Li
- abstract@[open-review](#): Recent studies show that training deep neural networks (DNNs) with Lipschitz constraints are able to enhance adversarial robustness and other model properties such as stability. In this paper, we propose a layer-wise orthogonal training method (LOT) to effectively train 1-Lipschitz convolution layers via parametrizing an orthogonal matrix with an unconstrained matrix. We then efficiently compute the inverse square root of a convolution kernel by transforming the input domain to the Fourier frequency domain. On the other hand, as existing works show that semi-supervised training helps improve empirical robustness, we aim to bridge the gap and prove that semi-supervised learning also improves the certified robustness of Lipschitz-bounded models. We conduct comprehensive evaluations for LOT under different settings. We show that LOT significantly outperforms baselines regarding deterministic L2 certified robustness, and scales to deeper neural networks. Under the supervised scenario, we improve the state-of-the-art certified robustness for all architectures (e.g. from 59.04% to 63.50% on CIFAR-10 and from 32.57% to 34.59% on CIFAR-100 at radius  $\$r\rho=36/255\$$  for 40-layer networks). With semi-supervised learning over unlabelled data, we are able to improve state-of-the-art certified robustness on CIFAR-10 at  $\$r\rho=108/255\$$  from 36.04% to 42.39%. In addition, LOT consistently outperforms baselines on different model architectures with only 1/3 evaluation time.

## [CoPur: Certifiably Robust Collaborative Inference via Feature Purification](#)

- Jing Liu · Chulin Xie · Sanmi Koyejo · Bo Li
- abstract@[open-review](#): Collaborative inference leverages diverse features provided by different agents (e.g., sensors) for more accurate inference. common setup is where each agent sends its embedded features instead of the raw data to the Fusion Center (FC) for joint prediction. In this setting, we consider the inference-time attacks when a small fraction of agents are compromised. The compromised agent either does not send embedded features to the FC, or sends arbitrarily embedded features. To address this, we propose a certifiably robust COllaborative inference framework via feature PURification (CoPur), by leveraging the block-sparse nature of adversarial perturbations on the feature vector, as well as exploring the underlying redundancy across the embedded features (by assuming the overall features lie on an underlying lower dimensional manifold). We theoretically show that the proposed feature purification method can robustly recover the true feature vector, despite adversarial corruptions and/or incomplete observations. We also propose and test an untargeted distributed feature-flipping attack, which is agnostic to the model, training data, label, as well as the features held by other agents, and is shown to be effective in attacking state-of-the-art defenses. Experiments on ExtraSensory and NUS-WIDE datasets show that CoPur significantly outperforms existing defenses in terms of robustness against targeted and untargeted adversarial attacks.

## [A Consolidated Cross-Validation Algorithm for Support Vector Machines via Data Reduction](#)

- Boxiang Wang · Archer Yang

- abstract@[open-review](#): We propose a consolidated cross-validation (CV) algorithm for training and tuning the support vector machines (SVM) on reproducing kernel Hilbert spaces. Our consolidated CV algorithm utilizes a recently proposed exact leave-one-out formula for the SVM and accelerates the SVM computation via a data reduction strategy. In addition, to compute the SVM with the bias term (intercept), which is not handled by the existing data reduction methods, we propose a novel two-stage consolidated CV algorithm. With numerical studies, we demonstrate that our algorithm is about an order of magnitude faster than the two mainstream SVM solvers, kernlab and LIBSVM, with almost the same accuracy.

## [Byzantine-tolerant federated Gaussian process regression for streaming data](#)

- Xu Zhang Â· Zhenyuan Yuan Â· Minghui Zhu
- abstract@[open-review](#): In this paper, we consider Byzantine-tolerant federated learning for streaming data using Gaussian process regression (GPR). In particular, a cloud and a group of agents aim to collaboratively learn a latent function where some agents are subject to Byzantine attacks. We develop a Byzantine-tolerant federated GPR algorithm, which includes three modules: agent-based local GPR, cloud-based aggregated GPR and agent-based fused GPR. Specifically, the agent-based local GPR sends potentially compromised local predictions to the cloud, and the cloud-based aggregated GPR computes a global model by a Byzantine-tolerant Product-of-Experts aggregation rule. Then the cloud broadcasts the global model to all the agents. Agent-based fused GPR refines the predictions by fusing the model from the cloud-based GPR with that from the agent-based local GPR. We derive the upper bounds on prediction error between the mean from the cloud-based aggregated GPR and the target function provided that Byzantine agents are less than one quarter of all the agents. We also characterize the lower and upper bounds of the predictive variance. Experiments on a synthetic dataset and two real-world datasets are conducted to evaluate the proposed algorithm.

## [Provable Subspace Identification Under Post-Nonlinear Mixtures](#)

- Qi Lyu Â· Xiao Fu
- abstract@[open-review](#): Unsupervised mixture learning (UML) aims at identifying linearly or nonlinearly mixed latent components in a blind manner. UML is known to be challenging: Even learning linear mixtures requires highly nontrivial analytical tools, e.g., independent component analysis or nonnegative matrix factorization. In this work, the post-nonlinear (PNL) mixture model is revisited, where {it unknown} element-wise nonlinear functions are imposed after a linear mixture, making the identification problem even more ill-posed. The PNL model is widely employed in different fields ranging from brain signal classification, speech separation, remote sensing, to causal discovery. Existing works often assume different properties on the latent components (e.g., statistical independence or probability-simplex structures) to identify and remove the unknown nonlinear functions. This work shows that the existence of a nontrivial {it null space} associated with the underlying mixing system suffices to guarantee identification/removal of the unknown nonlinearity. Compared to the existing works, this finding largely relaxes the model identification conditions of PNL models. Consequently, a simple learning criterion is proposed that could benefit applications where no strong structural information on the latent components is known. A finite-sample analysis is offered to characterize the performance of the proposed approach under realistic settings. For implementation, we design an optimization strategy that features a block coordinate descent algorithm. A series of numerical experiments corroborate our theoretical claims.

## [CoupAlign: Coupling Word-Pixel with Sentence-Mask Alignments for Referring Image Segmentation](#)

- Zicheng Zhang Â· Yi Zhu Â· Jianzhuang Liu Â· Xiaodan Liang Â· Wei Ke
- abstract@[open-review](#): Referring image segmentation aims at localizing all pixels of the visual objects described by a natural language sentence. Previous works learn to straightforwardly align the sentence embedding and pixel-level embedding for highlighting the referred objects, but ignore the semantic consistency of pixels within the same object, leading to incomplete masks and localization errors in predictions. To tackle this problem, we propose CoupAlign, a simple yet effective multi-level visual-semantic alignment method, to couple sentence-mask alignment with word-pixel alignment to enforce object mask constraint for achieving more accurate localization and segmentation. Specifically, the Word-Pixel Alignment (WPA) module performs early fusion of linguistic and pixel-level features in intermediate layers of the vision and language encoders. Based on the word-pixel aligned embedding, a set of mask proposals are generated to hypothesize possible objects. Then in the Sentence-Mask Alignment (SMA) module, the masks are weighted by the sentence embedding to localize the referred object, and finally projected back to aggregate the pixels for the target. To further enhance the learning of the two alignment modules, an auxiliary loss is designed to contrast the foreground and background pixels. By hierarchically aligning pixels and masks with linguistic features, our CoupAlign captures the pixel coherence at both visual and semantic levels, thus generating more accurate predictions. Extensive experiments on popular datasets (e.g., RefCOCO and G-Ref) show that our method achieves consistent improvements over state-of-the-art methods, e.g., about 2% oIoU increase on the validation and testing set of RefCOCO. Especially, CoupAlign has remarkable ability in distinguishing the target from multiple objects of the same class.

## [Towards Reasonable Budget Allocation in Untargeted Graph Structure Attacks via Gradient Debias](#)

- Zihan Liu Â· Yun Luo Â· Lirong Wu Â· Zicheng Liu Â· Stan Z. Li
- abstract@[open-review](#): It has become cognitive inertia to employ cross-entropy loss function in classification related tasks. In the untargeted attacks on graph structure, the gradients derived from the attack objective are the attacker's basis for evaluating a perturbation scheme. Previous methods use negative cross-entropy loss as the attack objective in attacking node-level classification models. However, the suitability of the cross-entropy function for constructing the untargeted attack objective has yet been discussed in previous works. This paper argues about the previous unreasonable attack objective from the perspective of budget allocation. We demonstrate theoretically and empirically that negative cross-entropy tends to produce more significant gradients from nodes with lower confidence in the labeled classes, even if the predicted classes of these nodes have been misled. To free up these inefficient attack budgets, we propose a simple attack model for untargeted attacks on graph structure based on a novel attack objective which generates unweighted gradients on graph structures that are not affected by the node confidence. By conducting experiments in gray-box poisoning attack scenarios, we demonstrate that a reasonable budget allocation can significantly improve the effectiveness of gradient-based edge perturbations without any extra hyper-parameter.

## [Enhancing Safe Exploration Using Safety State Augmentation](#)

- Aivar Sootla Â· Alexander Cowen-Rivers Â· Jun Wang Â· Haitham Bou Ammar
- abstract@[open-review](#): Safe exploration is a challenging and important problem in model-free reinforcement learning (RL). Often the safety cost is sparse and unknown, which unavoidably leads to constraint violations - a phenomenon ideally to be avoided in safety-critical applications. We tackle this problem by augmenting the state-space with a safety state, which is nonnegative if and only if the constraint is satisfied. The value of this state also serves as a distance toward constraint violation, while its initial value indicates the available safety budget. This idea allows us to derive policies for scheduling the safety budget during training. We call our approach Simmer (Safe policy IMproveMEnt for RL) to reflect the careful nature of these schedules. We apply this idea to two safe RL problems: RL with constraints imposed on an average cost, and RL with constraints imposed on a cost with probability one. Our experiments suggest that simmering a safe algorithm can improve safety during training for both settings. We further show that Simmer can stabilize training and improve the performance of safe RL with average constraints.

## [CS-Shapley: Class-wise Shapley Values for Data Valuation in Classification](#)

- Stephanie Schoch Â· Haifeng Xu Â· Yangfeng Ji
- abstract@[open-review](#): Data valuation, or the valuation of individual datum contributions, has seen growing interest in machine learning due to its demonstrable efficacy for tasks such as noisy label detection. In particular, due to the desirable axiomatic properties, several Shapley value approximations

have been proposed. In these methods, the value function is usually defined as the predictive accuracy over the entire development set. However, this limits the ability to differentiate between training instances that are helpful or harmful to their own classes. Intuitively, instances that harm their own classes may be noisy or mislabeled, and should be valued lower than instances that are helpful. In this work, we propose CS-Shapley, a Shapley value with a new value function that discriminates between training instances in-class and out-of-class contributions. Our theoretical analysis shows the proposed value function is (essentially) the unique function that satisfies two desirable properties for evaluating data values in classification. Further, our experiments on two benchmark evaluation tasks (data removal and noisy label detection) and four classifiers demonstrate the effectiveness of CS-Shapley over existing methods. Lastly, we evaluate the transferability of data values estimated from one classifier to others, and our results suggest Shapley-based data valuation is transferable for application across different models.

## [A Unifying Framework for Online Optimization with Long-Term Constraints](#)

- Matteo Castiglioni · Andrea Celli · Alberto Marchesi · Giulia Romano · Nicola Gatti
- abstract@[open-review](#): We study online learning problems in which a decision maker has to take a sequence of decisions subject to long-term constraints. The goal of the decision maker is to maximize their total reward, while at the same time achieving small cumulative constraints violations across the rounds. We present the first best-of-both-world type algorithm for this general class of problems, with no-regret guarantees both in the case in which rewards and constraints are selected according to an unknown stochastic model, and in the case in which they are selected at each round by an adversary. Our algorithm is the first to provide guarantees in the adversarial setting with respect to the optimal fixed strategy that satisfies the long-term constraints. In particular, it guarantees a  $\frac{\rho}{1+\rho}$  fraction of the optimal utility and sublinear regret, where  $\rho$  is a feasibility parameter related to the existence of strictly feasible solutions. Our framework employs traditional regret minimizers as black-box components. Therefore, by instantiating it with an appropriate choice of regret minimizers it can handle both the full-feedback as well as the bandit-feedback setting. Moreover, it allows the decision maker to seamlessly handle scenarios with non-convex reward and constraints. We show how our framework may be applied in the context of budget-management mechanisms for repeated auctions in order to guarantee long-term constraints which are not packing (e.g., ROI constraints).

## [Fast Bayesian Inference with Batch Bayesian Quadrature via Kernel Recombination](#)

- Masaki Adachi · Satoshi Hayakawa · Martin Jørgensen · Harald Oberhauser · Michael A Osborne
- abstract@[open-review](#): Calculation of Bayesian posteriors and model evidences typically requires numerical integration. Bayesian quadrature (BQ), a surrogate-model-based approach to numerical integration, is capable of superb sample efficiency, but its lack of parallelisation has hindered its practical applications. In this work, we propose a parallelised (batch) BQ method, employing techniques from kernel quadrature, that possesses an empirically exponential convergence rate. Additionally, just as with Nested Sampling, our method permits simultaneous inference of both posteriors and model evidence. Samples from our BQ surrogate model are re-selected to give a sparse set of samples, via a kernel recombination algorithm, requiring negligible additional time to increase the batch size. Empirically, we find that our approach significantly outperforms the sampling efficiency of both state-of-the-art BQ techniques and Nested Sampling in various real-world datasets, including lithium-ion battery analytics.

## [Lost in Latent Space: Examining failures of disentangled models at combinatorial generalisation](#)

- Milton Montero · Jeffrey Bowers · Rui Ponte Costa · Casimir Ludwig · Gaurav Malhotra
- abstract@[open-review](#): Recent research has shown that generative models with highly disentangled representations fail to generalise to unseen combination of generative factor values. These findings contradict earlier research which showed improved performance in out-of-training distribution settings when compared to entangled representations. Additionally, it is not clear if the reported failures are due to (a) encoders failing to map novel combinations to the proper regions of the latent space, or (b) novel combinations being mapped correctly but the decoder is unable to render the correct output for the unseen combinations. We investigate these alternatives by testing several models on a range of datasets and training settings. We find that (i) when models fail, their encoders also fail to map unseen combinations to correct regions of the latent space and (ii) when models succeed, it is either because the test conditions do not exclude enough examples, or because excluded cases involve combinations of object properties with its shape. We argue that to generalise properly, models not only need to capture factors of variation, but also understand how to invert the process that causes the visual stimulus.

## [Distributed Optimization for Overparameterized Problems: Achieving Optimal Dimension Independent Communication Complexity](#)

- Bingqing Song · Ioannis Tsaknakis · Chung-Yiu Yau · Hoi-To Wai · Mingyi Hong
- abstract@[open-review](#): Decentralized optimization are playing an important role in applications such as training large machine learning models, among others. Despite its superior practical performance, there has been some lack of fundamental understanding about its theoretical properties. In this work, we address the following open research question: To train an overparameterized model over a set of distributed nodes, what is the minimum communication overhead (in terms of the bits got exchanged) that the system needs to sustain, while still achieving (near) zero training loss? We show that for a class of overparameterized models where the number of parameters  $D$  is much larger than the total data samples  $N$ , the best possible communication complexity is  $\Omega(N)$ , which is independent of the problem dimension  $D$ . Further, for a few specific overparameterized models (i.e., the linear regression, and certain multi-layer neural network with one wide layer), we develop a set of algorithms which uses certain linear compression followed by adaptive quantization, and show that they achieve dimension independent, and sometimes near optimal, communication complexity. To our knowledge, this is the first time that dimension independent communication complexity has been shown for distributed optimization.

## [DIMES: A Differentiable Meta Solver for Combinatorial Optimization Problems](#)

- Ruizhong Qiu · Zhiqing Sun · Yiming Yang
- abstract@[open-review](#): Recently, deep reinforcement learning (DRL) models have shown promising results in solving NP-hard Combinatorial Optimization (CO) problems. However, most DRL solvers can only scale to a few hundreds of nodes for combinatorial optimization problems on graphs, such as the Traveling Salesman Problem (TSP). This paper addresses the scalability challenge in large-scale combinatorial optimization by proposing a novel approach, namely, DIMES. Unlike previous DRL methods which suffer from costly autoregressive decoding or iterative refinements of discrete solutions, DIMES introduces a compact continuous space for parameterizing the underlying distribution of candidate solutions. Such a continuous space allows stable REINFORCE-based training and fine-tuning via massively parallel sampling. We further propose a meta-learning framework to enable the effective initialization of model parameters in the fine-tuning stage. Extensive experiments show that DIMES outperforms recent DRL-based methods on large benchmark datasets for Traveling Salesman Problems and Maximal Independent Set problems.

## [Learning Debiased Classifier with Biased Committee](#)

- Nayeong Kim · SEHYUN HWANG · Sungsoo Ahn · Jaesik Park · Suha Kwak
- abstract@[open-review](#): Neural networks are prone to be biased towards spurious correlations between classes and latent attributes exhibited in a major portion of training data, which ruins their generalization capability. This paper proposes a new method for training debiased classifiers with no spurious attribute label. The key idea of the method is to employ a committee of classifiers as an auxiliary module that identifies bias-conflicting data, i.e., data without spurious correlations, and assigns large weights to them when training the main classifier. The committee is learned as a bootstrapped ensemble so that a majority of its classifiers are biased as well as being diverse, and intentionally fail to predict classes of bias-conflicting data accordingly. The

consensus within the committee on prediction difficulty thus provides a reliable cue for identifying and weighting bias-conflicting data. Moreover, the committee is also trained with knowledge transferred from the main classifier so that it gradually becomes debiased along with the target and emphasizes more difficult data as training progresses. On five real-world datasets, our method outperforms existing methods using no spurious attribute label like ours and even surpasses those relying on bias label occasionally.

## [Weakly supervised causal representation learning](#)

- Johann Brehmer · Pim de Haan · Phillip Lippe · Taco Cohen
- abstract@[open-review](#): Learning high-level causal representations together with a causal model from unstructured low-level data such as pixels is impossible from observational data alone. We prove under mild assumptions that this representation is however identifiable in a weakly supervised setting. This involves a dataset with paired samples before and after random, unknown interventions, but no further labels. We then introduce implicit latent causal models, variational autoencoders that represent causal variables and causal structure without having to optimize an explicit discrete graph structure. On simple image data, including a novel dataset of simulated robotic manipulation, we demonstrate that such models can reliably identify the causal structure and disentangle causal variables.

## [Generalization Bounds for Equivariant Networks](#)

- Arash Behboodi · Gabriele Cesa · Taco Cohen
- abstract@[open-review](#): Equivariant networks capture the inductive bias about the symmetry of the learning task by building those symmetries into the model. In this paper, we study how equivariance relates to generalization error utilizing PAC Bayesian analysis for equivariant networks, where the transformation laws of feature spaces are determined by group representations. By using perturbation analysis of equivariant networks in Fourier domain for each layer, we derive norm-based PAC-Bayesian generalization bounds. The bound characterizes the impact of group size, and multiplicity and degree of irreducible representations on the generalization error and thereby provide a guideline for selecting them. In general, the bound indicates that using larger group size in the model improves the generalization error substantiated by extensive numerical experiments.

## [Maximum-Likelihood Inverse Reinforcement Learning with Finite-Time Guarantees](#)

- Siliang Zeng · Chenliang Li · Alfredo Garcia · Mingyi Hong
- abstract@[open-review](#): Inverse reinforcement learning (IRL) aims to recover the reward function and the associated optimal policy that best fits observed sequences of states and actions implemented by an expert. Many algorithms for IRL have an inherent nested structure: the inner loop finds the optimal policy given parametrized rewards while the outer loop updates the estimates towards optimizing a measure of fit. For high dimensional environments such nested-loop structure entails a significant computational burden. To reduce the computational burden of a nested loop, novel methods such as SQIL \cite{reddy2019sql} and IQ-Learn \cite{garg2021iq} emphasize policy estimation at the expense of reward estimation accuracy. However, without accurate estimated rewards, it is not possible to do counterfactual analysis such as predicting the optimal policy under different environment dynamics and/or learning new tasks. In this paper we develop a novel \emph{single-loop} algorithm for IRL that does not compromise reward estimation accuracy. In the proposed algorithm, each policy improvement step is followed by a stochastic gradient step for likelihood maximization. We show that the proposed algorithm provably converges to a stationary solution with a finite-time guarantee. If the reward is parameterized linearly we show the identified solution corresponds to the solution of the maximum entropy IRL problem. Finally, by using robotics control problems in Mujoco and their transfer settings, we show that the proposed algorithm achieves superior performance compared with other IRL and imitation learning benchmarks.

## [MultiScan: Scalable RGBD scanning for 3D environments with articulated objects](#)

- Yongsen Mao · Yiming Zhang · Hanxiao Jiang · Angel Chang · Manolis Savva
- abstract@[open-review](#): We introduce MultiScan, a scalable RGBD dataset construction pipeline leveraging commodity mobile devices to scan indoor scenes with articulated objects and web-based semantic annotation interfaces to efficiently annotate object and part semantics and part mobility parameters. We use this pipeline to collect 230 scans of 108 indoor scenes containing 9458 objects and 4331 parts. The resulting MultiScan dataset provides RGBD streams with per-frame camera poses, textured 3D surface meshes, richly annotated part-level and object-level semantic labels, and part mobility parameters. We validate our dataset on instance segmentation and part mobility estimation tasks and benchmark methods for these tasks from prior work. Our experiments show that part segmentation and mobility estimation in real 3D scenes remain challenging despite recent progress in 3D object segmentation.

## [Hierarchical Lattice Layer for Partially Monotone Neural Networks](#)

- Hiroki Yanagisawa · Kohei Miyaguchi · Takayuki Katsuki
- abstract@[open-review](#): Partially monotone regression is a regression analysis in which the target values are monotonically increasing with respect to a subset of input features. The TensorFlow Lattice library is one of the standard machine learning libraries for partially monotone regression. It consists of several neural network layers, and its core component is the lattice layer. One of the problems of the lattice layer is its requirement for a special training algorithm to satisfy monotonicity constraints. Another problem is that it cannot receive a high-dimensional input vector due to the resultant memory consumption. We propose a novel neural network layer, the hierarchical lattice layer (HLL), as an extension of the lattice layer so that we can use a standard neural network algorithm to train HLL while satisfying monotonicity constraints and so that it can receive a high-dimensional input vector. Our experiments demonstrate that HLL did not sacrifice its prediction performance on real datasets compared with the lattice layer.

## [Batch Multi-Fidelity Active Learning with Budget Constraints](#)

- Shibo Li · Jeff M Phillips · Xin Yu · Robert Kirby · Shandian Zhe
- abstract@[open-review](#): Learning functions with high-dimensional outputs is critical in many applications, such as physical simulation and engineering design. However, collecting training examples for these applications is often costly, e.g., by running numerical solvers. The recent work (Li et al., 2022) proposes the first multi-fidelity active learning approach for high-dimensional outputs, which can acquire examples at different fidelities to reduce the cost while improving the learning performance. However, this method only queries at one pair of fidelity and input at a time, and hence has a risk to bring in strongly correlated examples to reduce the learning efficiency. In this paper, we propose Batch Multi-Fidelity Active Learning with Budget Constraints (BMFAL-BC), which can promote the diversity of training examples to improve the benefit-cost ratio, while respecting a given budget constraint for batch queries. Hence, our method can be more practically useful. Specifically, we propose a novel batch acquisition function that measures the mutual information between a batch of multi-fidelity queries and the target function, so as to penalizes highly correlated queries and encourages diversity. The optimization of the batch acquisition function is challenging in that it involves a combinatorial search over many fidelities while subject to the budget constraint. To address this challenge, we develop a weighted greedy algorithm that can sequentially identify each (fidelity, input) pair, while achieving a near  $\$(1 - 1/e)$ -approximation of the optimum. We show the advantage of our method in several computational physics and engineering applications.

## [MORA: Improving Ensemble Robustness Evaluation with Model Reweighting Attack](#)

- yunrui yu · Xitong Gao · Cheng-Zhong Xu

- abstract@[open-review](#): Adversarial attacks can deceive neural networks by adding tiny perturbations to their input data. Ensemble defenses, which are trained to minimize attack transferability among sub-models, offer a promising research direction to improve robustness against such attacks while maintaining a high accuracy on natural inputs. We discovered, however, that recent state-of-the-art (SOTA) adversarial attack strategies cannot reliably evaluate ensemble defenses, sizeably overestimating their robustness. This paper identifies the two factors that contribute to this behavior. First, these defenses form ensembles that are notably difficult for existing gradient-based method to attack, due to gradient obfuscation. Second, ensemble defenses diversify sub-model gradients, presenting a challenge to defeat all sub-models simultaneously, simply summing their contributions may counteract the overall attack objective. Yet, we observed that ensemble can still be fooled despite most sub-models being correct. We therefore introduce MORA, a model-reweighing attack to steer adversarial example synthesis by reweighing the importance of sub-model gradients. MORA discovered that recent ensemble defenses all exhibit varying degrees of overestimated robustness. Comparing it against recent SOTA white-box attacks, it can converge orders of magnitude faster while achieving higher attack success rates across all ensemble models examined with three different ensemble modes (i.e. forming ensembles by either softmax, voting or logits). In particular, most ensemble defenses exhibit near or exactly  $\$0\%$  robustness against MORA with  $\$ell^\infty$  perturbation within \$0.02\$ on CIFAR-10, and \$0.01\$ on CIFAR-100.

## [Optimistic Posterior Sampling for Reinforcement Learning with Few Samples and Tight Guarantees](#)

- Daniil Tiapkin · Denis Belomestny · Daniele Calandriello · Eric Moulines · Remi Munos · Alexey Naumov · Mark Rowland · Michal Valko · Pierre Ménard
- abstract@[open-review](#): We consider reinforcement learning in an environment modeled by an episodic, tabular, step-dependent Markov decision process of horizon  $H$  with  $S$  states, and  $A$  actions. The performance of an agent is measured by the regret after interacting with the environment for  $T$  episodes. We propose an optimistic posterior sampling algorithm for reinforcement learning (OPSRL), a simple variant of posterior sampling that only needs a number of posterior samples logarithmic in  $H$ ,  $S$ ,  $A$ , and  $T$  per state-action pair. For OPSRL we guarantee a high-probability regret bound of order at most  $O(\sqrt{H^3SAT})$  ignoring  $\text{poly}(\log(HSAT))$  terms. The key novel technical ingredient is a new sharp anti-concentration inequality for linear forms of a Dirichlet random vector which may be of independent interest. Specifically, we extend the normal approximation-based lower bound for Beta distributions by Alfors and Dinges (1984) to Dirichlet distributions. Our bound matches the lower bound of order  $\Omega(\sqrt{H^3SAT})$ , thereby answering the open problems raised by Agrawal and Jia (2017) for the episodic setting.

## [Tensor Program Optimization with Probabilistic Programs](#)

- Junru Shao · Xiyou Zhou · Siyuan Feng · Bohan Hou · Ruihang Lai · Hongyi Jin · Wuwei Lin · Masahiro Masuda · Cody Hao Yu · Tianqi Chen
- abstract@[open-review](#): Automatic optimization for tensor programs becomes increasingly important as we deploy deep learning in various environments, and efficient optimization relies on a rich search space and effective search. Most existing efforts adopt a search space which lacks the ability to efficiently enable domain experts to grow the search space. This paper introduces SpaceCraft, a domain-specific probabilistic programming language abstraction to construct a rich search space of tensor programs. Our abstraction allows domain experts to analyze the program, and easily propose stochastic choices in a modular way to compose program transformation accordingly. We also build an end-to-end learning-driven framework to find an optimized program for a given search space. Experimental results show that SpaceCraft can cover the search space used in the state-of-the-art tensor program optimization frameworks in a modular way. Additionally, it empowers domain experts to conveniently grow the search space and modularly enhance the system, which brings 48% speedup on end-to-end deep learning workloads.

## [FOF: Learning Fourier Occupancy Field for Monocular Real-time Human Reconstruction](#)

- Qiao Feng · Yebin Liu · Yu-Kun Lai · Jingyu Yang · Kun Li
- abstract@[open-review](#): The advent of deep learning has led to significant progress in monocular human reconstruction. However, existing representations, such as parametric models, voxel grids, meshes and implicit neural representations, have difficulties achieving high-quality results and real-time speed at the same time. In this paper, we propose Fourier Occupancy Field (FOF), a novel powerful, efficient and flexible 3D representation, for monocular real-time and accurate human reconstruction. The FOF represents a 3D object with a 2D field orthogonal to the view direction where at each 2D position the occupancy field of the object along the view direction is compactly represented with the first few terms of Fourier series, which retains the topology and neighborhood relation in the 2D domain. A FOF can be stored as a multi-channel image, which is compatible with 2D convolutional neural networks and can bridge the gap between 3D geometries and 2D images. The FOF is very flexible and extensible, e.g., parametric models can be easily integrated into a FOF as a prior to generate more robust results. Based on FOF, we design the first 30+FPS high-fidelity real-time monocular human reconstruction framework. We demonstrate the potential of FOF on both public dataset and real captured data. The code will be released for research purposes.

## [PAC-Bayes Compression Bounds So Tight That They Can Explain Generalization](#)

- Sanae Lotfi · Sanyam Kapoor · Marc Finzi · Andres Potapczynski · Micah Goldblum · Andrew Wilson
- abstract@[open-review](#): While there has been progress in developing non-vacuous generalization bounds for deep neural networks, these bounds tend to be uninformative about why deep learning works. In this paper, we develop a compression approach based on quantizing neural network parameters in a linear subspace, profoundly improving on previous results to provide state-of-the-art generalization bounds on a variety of tasks, including transfer learning. We use these tight bounds to better understand the role of model size, equivariance, and the implicit biases of optimization, for generalization in deep learning. Notably, we find large models can be compressed to a much greater extent than previously known, encapsulating Occam's razor.

## [Set-based Meta-Interpolation for Few-Task Meta-Learning](#)

- Seanie Lee · Bruno Andreis · Kenji Kawaguchi · Juho Lee · Sung Ju Hwang
- abstract@[open-review](#): Meta-learning approaches enable machine learning systems to adapt to new tasks given few examples by leveraging knowledge from related tasks. However, a large number of meta-training tasks are still required for generalization to unseen tasks during meta-testing, which introduces a critical bottleneck for real-world problems that come with only few tasks, due to various reasons including the difficulty and cost of constructing tasks. Recently, several task augmentation methods have been proposed to tackle this issue using domain-specific knowledge to design augmentation techniques to densify the meta-training task distribution. However, such reliance on domain-specific knowledge renders these methods inapplicable to other domains. While Manifold Mixup based task augmentation methods are domain-agnostic, we empirically find them ineffective on non-image domains. To tackle these limitations, we propose a novel domain-agnostic task augmentation method, Meta-Interpolation, which utilizes expressive neural set functions to densify the meta-training task distribution using bilevel optimization. We empirically validate the efficacy of Meta-Interpolation on eight datasets spanning across various domains such as image classification, molecule property prediction, text classification and speech recognition. Experimentally, we show that Meta-Interpolation consistently outperforms all the relevant baselines. Theoretically, we prove that task interpolation with the set function regularizes the meta-learner to improve generalization. We provide our source code in the supplementary material.

## [In Differential Privacy, There is Truth: on Vote-Histogram Leakage in Ensemble Private Learning](#)

- JIAQI WANG · Roei Schuster · I Shumailov · David Lie · Nicolas Papernot
- abstract@[open-review](#): When learning from sensitive data, care must be taken to ensure that training algorithms address privacy concerns. The canonical Private Aggregation of Teacher Ensembles, or PATE, computes output labels by aggregating the predictions of a (possibly distributed) collection of teacher models via a voting mechanism. The mechanism adds noise to attain a differential privacy guarantee with respect to the teachers' training data. In

this work, we observe that this use of noise, which makes PATE predictions stochastic, enables new forms of leakage of sensitive information. For a given input, our adversary exploits this stochasticity to extract high-fidelity histograms of the votes submitted by the underlying teachers. From these histograms, the adversary can learn sensitive attributes of the input such as race, gender, or age. Although this attack does not directly violate the differential privacy guarantee, it clearly violates privacy norms and expectations, and would not be possible without the noise inserted to obtain differential privacy. In fact, counter-intuitively, the attack becomes easier as we add more noise to provide stronger differential privacy. We hope this encourages future work to consider privacy holistically rather than treat differential privacy as a panacea.

## [Unsupervised Learning of Group Invariant and Equivariant Representations](#)

- Robin Winter · Marco Bertolini · Tuan Le · Frank Noe · Djork-Arné Clevert
- abstract@[open-review](#): Equivariant neural networks, whose hidden features transform according to representations of a group  $G$  acting on the data, exhibit training efficiency and an improved generalisation performance. In this work, we extend group invariant and equivariant representation learning to the field of unsupervised deep learning. We propose a general learning strategy based on an encoder-decoder framework in which the latent representation is separated in an invariant term and an equivariant group action component. The key idea is that the network learns to encode and decode data to and from a group-invariant representation by additionally learning to predict the appropriate group action to align input and output pose to solve the reconstruction task. We derive the necessary conditions on the equivariant encoder, and we present a construction valid for any  $G$ , both discrete and continuous. We describe explicitly our construction for rotations, translations and permutations. We test the validity and the robustness of our approach in a variety of experiments with diverse data types employing different network architectures.

## [Provably Efficient Offline Multi-agent Reinforcement Learning via Strategy-wise Bonus](#)

- Qiwen Cui · Simon Du
- abstract@[open-review](#): This paper considers offline multi-agent reinforcement learning. We propose the strategy-wise concentration principle which directly builds a confidence interval for the joint strategy, in contrast to the point-wise concentration principle which builds a confidence interval for each point in the joint action space. For two-player zero-sum Markov games, by exploiting the convexity of the strategy-wise bonus, we propose a computationally efficient algorithm whose sample complexity enjoys a better dependency on the number of actions than the prior methods based on the point-wise bonus. Furthermore, for offline multi-agent general-sum Markov games, based on the strategy-wise bonus and a novel surrogate function, we give the first algorithm whose sample complexity only scales  $\sum_{i=1}^m A_i$  where  $A_i$  is the action size of the  $i$ -th player and  $m$  is the number of players. In sharp contrast, the sample complexity of methods based on the point-wise bonus would scale with the size of the joint action space  $|Pi|$  due to the curse of multiagents. Lastly, all of our algorithms can naturally take a pre-specified strategy class  $Pi$  as input and output a strategy that is close to the best strategy in  $Pi$ . In this setting, the sample complexity only scales with  $\log |Pi|$  instead of  $\sum_{i=1}^m A_i$ .

## [Practical Adversarial Attacks on Spatiotemporal Traffic Forecasting Models](#)

- Fan LIU · Hao Liu · Wenzhao Jiang
- abstract@[open-review](#): Machine learning based traffic forecasting models leverage sophisticated spatiotemporal auto-correlations to provide accurate predictions of city-wide traffic states. However, existing methods assume a reliable and unbiased forecasting environment, which is not always available in the wild. In this work, we investigate the vulnerability of spatiotemporal traffic forecasting models and propose a practical adversarial spatiotemporal attack framework. Specifically, instead of simultaneously attacking all geo-distributed data sources, an iterative gradient guided node saliency method is proposed to identify the time-dependent set of victim nodes. Furthermore, we devise a spatiotemporal gradient descent based scheme to generate real-valued adversarial traffic states under a perturbation constraint. Meanwhile, we theoretically demonstrate the worst performance bound of adversarial traffic forecasting attacks. Extensive experiments on two real-world datasets show that the proposed two-step framework achieves up to 67.8% performance degradation on various advanced spatiotemporal forecasting models. Remarkably, we also show that adversarial training with our proposed attacks can significantly improve the robustness of spatiotemporal traffic forecasting models.

## [Gradient flow dynamics of shallow ReLU networks for square loss and orthogonal inputs](#)

- Etienne Boursier · Loucas PILLAUD-VIVIEN · Nicolas Flammarion
- abstract@[open-review](#): The training of neural networks by gradient descent methods is a cornerstone of the deep learning revolution. Yet, despite some recent progress, a complete theory explaining its success is still missing. This article presents, for orthogonal input vectors, a precise description of the gradient flow dynamics of training one-hidden layer ReLU neural networks for the mean squared error at small initialisation. In this setting, despite non-convexity, we show that the gradient flow converges to zero loss and characterise its implicit bias towards minimum variation norm. Furthermore, some interesting phenomena are highlighted: a quantitative description of the initial alignment phenomenon and a proof that the process follows a specific saddle to saddle dynamics.

## [Energy-Based Contrastive Learning of Visual Representations](#)

- Beomsu Kim · Jong Chul Ye
- abstract@[open-review](#): Contrastive learning is a method of learning visual representations by training Deep Neural Networks (DNNs) to increase the similarity between representations of positive pairs (transformations of the same image) and reduce the similarity between representations of negative pairs (transformations of different images). Here we explore Energy-Based Contrastive Learning (EBCLR) that leverages the power of generative learning by combining contrastive learning with Energy-Based Models (EBMs). EBCLR can be theoretically interpreted as learning the joint distribution of positive pairs, and it shows promising results on small and medium-scale datasets such as MNIST, Fashion-MNIST, CIFAR-10, and CIFAR-100. Specifically, we find EBCLR demonstrates from  $\times 4$  up to  $\times 20$  acceleration compared to SimCLR and MoCo v2 in terms of training epochs. Furthermore, in contrast to SimCLR, we observe EBCLR achieves nearly the same performance with  $254$  negative pairs (batch size  $128$ ) and  $30$  negative pairs (batch size  $16$ ) per positive pair, demonstrating the robustness of EBCLR to small numbers of negative pairs. Hence, EBCLR provides a novel avenue for improving contrastive learning methods that usually require large datasets with a significant number of negative pairs per iteration to achieve reasonable performance on downstream tasks.

## [Learning Best Combination for Efficient N:M Sparsity](#)

- Yuxin Zhang · Mingbao Lin · ZhiHang Lin · Yiting Luo · Ke Li · Fei Chao · Yongjian Wu · Rongrong Ji
- abstract@[open-review](#): By forcing  $N$  out of  $M$  consecutive weights to be non-zero, the recent N:M fine-grained network sparsity has received increasing attention with its two attractive advantages over traditional irregular network sparsity methods: 1) Promising performance at a high sparsity. 2) Significant speedups when performed on NVIDIA A100 GPUs. Current implementation on N:M sparsity requires a tedious pre-training phase or computationally heavy from-scratch training. To circumvent these problems, this paper presents an efficient solution for achieving N:M fine-grained sparsity from scratch. Specifically, we first make a re-formulation to convert the N:M fine-grained sparsity into a combinatorial problem, in which, the object falls into choosing the best weight combination among  $C_M^N$  candidates. Then, we equip each combination with a learnable importance score, which can be jointly optimized along with its associated weights. Through rigorous proof, we demonstrate that the magnitude of the optimized score well reflects the importance of its corresponding weights combination to the training loss. Therefore, by gradually removing combinations with smaller scores till the best one is left, N:M fine-grained sparsity can be efficiently optimized during the normal training phase without any extra expenditure. Comprehensive

experimental results have demonstrated that our proposed method for learning best combination, dubbed as LBC, consistently increases the efficacy of the off-the-shelf N:M methods across varying networks and datasets. Our project is released at <https://github.com/zyxxmu/LBC>.

## [An Empirical Study on Disentanglement of Negative-free Contrastive Learning](#)

- Jinkun Cao · Ruiqian Nai · Qing Yang · Jialei Huang · Yang Gao
- abstract@[open-review](#): Negative-free contrastive learning has attracted a lot of attention with simplicity and impressive performances for large-scale pretraining. But its disentanglement property remains unexplored. In this paper, we take different negative-free contrastive learning methods to study the disentanglement property of this genre of self-supervised methods empirically. We find the existing disentanglement metrics fail to make meaningful measurements for the high-dimensional representation model so we propose a new disentanglement metric based on Mutual Information between representation and data factors. With the proposed metric, we benchmark the disentanglement property of negative-free contrastive learning for the first time, on both popular synthetic datasets and a real-world dataset CelebA. Our study shows that the investigated methods can learn a well-disentangled subset of representation. We extend the study of the disentangled representation learning to high-dimensional representation space and negative-free contrastive learning for the first time.

## [Merging Models with Fisher-Weighted Averaging](#)

- Michael S Matena · Colin Raffel
- abstract@[open-review](#): Averaging the parameters of models that have the same architecture and initialization can provide a means of combining their respective capabilities. In this paper, we take the perspective that this merging" operation can be seen as choosing parameters that approximately maximize the joint likelihood of the posteriors of the models' parameters. Computing a simple average of the models' parameters therefore corresponds to making an isotropic Gaussian approximation to their posteriors. We develop an alternative merging procedure based on the Laplace approximation where we approximate each model's posterior as a Gaussian distribution whose precision matrix corresponds to its Fisher information. We first show that ourFisher merging" technique provides a performance boost in settings where simple parameter averaging is currently used -- specifically, robust fine-tuning and model ensembling. Then, we compare merging to standard gradient-based transfer learning and demonstrate that merging enables a fundamentally different method for transferring capabilities across models. Specifically, we show that Fisher merging is competitive with gradient-based transfer learning approaches (while being significantly cheaper) in intermediate-task training and domain-adaptive pre-training. We also show that our merging procedure makes it possible to combine models in previously unexplored ways. We release our code to facilitate future research into methods for merging models.

## [Neural Conservation Laws: A Divergence-Free Perspective](#)

- Jack Richter-Powell · Yaron Lipman · Ricky T. Q. Chen
- abstract@[open-review](#): We investigate the parameterization of deep neural networks that are by design divergence-free. To motivate the importance of divergence-free models, we observe that the continuity equation, a fundamental conservation law, can be characterized through a divergence-free vector field. We hence propose an approach to building divergence-free neural networks through the concept of differential forms, and with the aid of automatic differentiation, realize two practical constructions with differing trade offs. Furthermore, we show these models are universal and can be used to represent any divergence-free vector field if given sufficient capacity. As a result, we can parameterize density and vector fields that always satisfy the continuity equation simply by design, foregoing the need for expensive numerical simulation. We experimentally validate our approaches on neural network-based solutions to fluid equations and on learning dynamical optimal transport maps.

## [Adapting Self-Supervised Vision Transformers by Probing Attention-Conditioned Masking Consistency](#)

- Viraj Prabhu · Sriram Yenamandra · Aaditya Singh · Judy Hoffman
- abstract@[open-review](#): Visual domain adaptation (DA) seeks to transfer trained models to unseen, unlabeled domains across distribution shift, but approaches typically focus on adapting convolutional neural network architectures initialized with supervised ImageNet representations. In this work, we shift focus to adapting modern architectures for object recognition -- the increasingly popular Vision Transformer (ViT) -- initialized with modern pretraining based on self-supervised learning (SSL). Inspired by the design of recent SSL approaches based on learning from partial image inputs generated via masking or cropping -- either by learning to predict the missing pixels, or learning representational invariances to such augmentations -- we propose PACMAC, a two-stage adaptation algorithm for self-supervised ViTs. PACMAC first performs in-domain SSL on pooled source and target data to learn task-discriminative features, and then probes the model's predictive consistency across a set of partial target inputs generated via a novel attention-conditioned masking strategy, to identify reliable candidates for self-training. Our simple approach leads to consistent performance gains over competing methods that use ViTs and self-supervised initializations on standard object recognition benchmarks.

## [Robust \\$\phi\\$-Divergence MDPs](#)

- Chin Pang Ho · Marek Petrik · Wolfram Wiesemann
- abstract@[open-review](#): In recent years, robust Markov decision processes (MDPs) have emerged as a prominent modeling framework for dynamic decision problems affected by uncertainty. In contrast to classical MDPs, which only account for stochasticity by modeling the dynamics through a stochastic process with a known transition kernel, a robust MDP additionally accounts for ambiguity by optimizing in view of the most adverse transition kernel from a prescribed ambiguity set. In this paper, we develop a novel solution framework for robust MDPs with  $\phi$ -rectangular ambiguity sets that decomposes the problem into a sequence of robust Bellman updates and simplex projections. Exploiting the rich structure present in the simplex projections corresponding to  $\phi$ -divergence ambiguity sets, we show that the associated  $\phi$ -rectangular robust MDPs can be solved substantially faster than with state-of-the-art commercial solvers as well as a recent first-order solution scheme, thus rendering them attractive alternatives to classical MDPs in practical applications.

## [On Kernelized Multi-Armed Bandits with Constraints](#)

- Xingyu Zhou · Bo Ji
- abstract@[open-review](#): We study a stochastic bandit problem with a general unknown reward function and a general unknown constraint function. Both functions can be non-linear (even non-convex) and are assumed to lie in a reproducing kernel Hilbert space (RKHS) with a bounded norm. This kernelized bandit setup strictly generalizes standard multi-armed bandits and linear bandits. In contrast to safety-type hard constraints studied in prior works, we consider soft constraints that may be violated in any round as long as the cumulative violations are small, which is motivated by various practical applications. Our ultimate goal is to study how to utilize the nature of soft constraints to attain a finer complexity-regret-constraint trade-off in the kernelized bandit setting. To this end, leveraging primal-dual optimization, we propose a general framework for both algorithm design and performance analysis. This framework builds upon a novel sufficient condition, which not only is satisfied under general exploration strategies, including upper confidence bound (UCB), Thompson sampling (TS), and new ones based on random exploration, but also enables a unified analysis for showing both sublinear regret and sublinear or even zero constraint violation. We demonstrate the superior performance of our proposed algorithms via numerical experiments based on both synthetic and real-world datasets. Along the way, we also make the first detailed comparison between two popular methods for analyzing constrained bandits and Markov decision processes (MDPs) by discussing the key difference and some subtleties in the analysis, which could be of independent interest to the communities.

## [First is Better Than Last for Language Data Influence](#)

- Chih-Kuan Yeh · Ankur Taly · Mukund Sundararajan · Frederick Liu · Pradeep Ravikumar
- abstract@[open-review](#): The ability to identify influential training examples enables us to debug training data and explain model behavior. Existing techniques to do so are based on the flow of training data influence through the model parameters. For large models in NLP applications, it is often computationally infeasible to study this flow through all model parameters, therefore techniques usually pick the last layer of weights. However, we observe that since the activation connected to the last layer of weights contains "shared logic", the data influenced calculated via the last layer weights prone to cancellation effect", where the data influence of different examples have large magnitude that contradicts each other. The cancellation effect lowers the discriminative power of the influence score, and deleting influential examples according to this measure often does not change the model's behavior by much. To mitigate this, we propose a technique called TracIn-WE that modifies a method called TracIn to operate on the word embedding layer instead of the last layer, where the cancellation effect is less severe. One potential concern is that influence based on the word embedding layer may not encode sufficient high level information. However, we find that gradients (unlike embeddings) do not suffer from this, possibly because they chain through higher layers. We show that TracIn-WE significantly outperforms other data influence methods applied on the last layer by \$4-10\times\$ on the case deletion evaluation on three language classification tasks. In addition, TracIn-WE can produce scores not just at the level of the overall training input, but also at the level of words within the training input, a further aid in debugging.

## [Spatially Sparse Inference for Deep Generative Image Editing](#)

- Muyang Li · Ji Lin · Chenlin Meng · Stefano Ermon · Song Han · Jun-Yan Zhu
- abstract@[open-review](#): Deep generative models excel at synthesizing photo-realistic images and enable various image synthesis and editing applications. However, when editing a photo, existing methods tend to re-synthesize the entire output from scratch, including the unedited region, leading to a significant waste of computation, especially for minor editing operations. In this work, we present Spatially Sparse Inference (SSI), a general-purpose speedup technique that selectively performs computation for edited regions and is compatible with different types of generative models. Our key observation is that user editing is often incremental in the interactive setting. This allows us to pre-compute the feature maps of the original image. Given an edited image, we sparsely apply the filters to the edited regions while reusing the pre-computed features for the unedited regions. Based on our algorithm, we propose Sparse Incremental Generative Engine (SIGE) to convert the theoretical computation reduction to latency reduction on commonly-used hardware. With 1.2% area edited regions, our method reduces the computation of DDIM by \$7.5\times\$ and GauGAN by \$18\times\$ while preserving the visual fidelity. With engineabbr, we accelerate the inference time of DDIM by \$3.0\times\$ on RTX3090 and \$6.6\times\$ on Apple M1 Pro, and GauGAN by \$4.2\times\$ on RTX3090 and \$14\times\$ on Apple M1 Pro.

## [Moderate-fitting as a Natural Backdoor Defender for Pre-trained Language Models](#)

- Biru Zhu · Yujia Qin · Ganqu Cui · Yangyi Chen · Weilin Zhao · Chong Fu · Yangdong Deng · Zhiyuan Liu · Jingang Wang · Wei Wu · Maosong Sun · Ming Gu
- abstract@[open-review](#): Despite the great success of pre-trained language models (PLMs) in a large set of natural language processing (NLP) tasks, there has been a growing concern about their security in real-world applications. Backdoor attack, which poisons a small number of training samples by inserting backdoor triggers, is a typical threat to security. Trained on the poisoned dataset, a victim model would perform normally on benign samples but predict the attacker-chosen label on samples containing pre-defined triggers. The vulnerability of PLMs under backdoor attacks has been proved with increasing evidence in the literature. In this paper, we present several simple yet effective training strategies that could effectively defend against such attacks. To the best of our knowledge, this is the first work to explore the possibility of backdoor-free adaptation for PLMs. Our motivation is based on the observation that, when trained on the poisoned dataset, the PLM's adaptation follows a strict order of two stages: (1) a moderate-fitting stage, where the model mainly learns the major features corresponding to the original task instead of subsidiary features of backdoor triggers, and (2) an overfitting stage, where both features are learned adequately. Therefore, if we could properly restrict the PLM's adaptation to the moderate-fitting stage, the model would neglect the backdoor triggers but still achieve satisfying performance on the original task. To this end, we design three methods to defend against backdoor attacks by reducing the model capacity, training epochs, and learning rate, respectively. Experimental results demonstrate the effectiveness of our methods in defending against several representative NLP backdoor attacks. We also perform visualization-based analysis to attain a deeper understanding of how the model learns different features, and explore the effect of the poisoning ratio. Finally, we explore whether our methods could defend against backdoor attacks for the pre-trained CV model.

## [SoLar: Sinkhorn Label Refinery for Imbalanced Partial-Label Learning](#)

- Haobo Wang · Mingxuan Xia · Yixuan Li · Yuren Mao · Lei Feng · Gang Chen · Junbo Zhao
- abstract@[open-review](#): Partial-label learning (PLL) is a peculiar weakly-supervised learning task where the training samples are generally associated with a set of candidate labels instead of single ground truth. While a variety of label disambiguation methods have been proposed in this domain, they normally assume a class-balanced scenario that may not hold in many real-world applications. Empirically, we observe degenerated performance of the prior methods when facing the combinatorial challenge from the long-tailed distribution and partial-labeling. In this work, we first identify the major reasons that the prior work failed. We subsequently propose SoLar, a novel Optimal Transport-based framework that allows to refine the disambiguated labels towards matching the marginal class prior distribution. SoLar additionally incorporates a new and systematic mechanism for estimating the long-tailed class prior distribution under the PLL setup. Through extensive experiments, SoLar exhibits substantially superior results on standardized benchmarks compared to the previous state-of-the-art PLL methods.

## [Constrained Stochastic Nonconvex Optimization with State-dependent Markov Data](#)

- Abhishek Roy · Krishnakumar Balasubramanian · Saeed Ghadimi
- abstract@[open-review](#): We study stochastic optimization algorithms for constrained nonconvex stochastic optimization problems with Markovian data. In particular, we focus on the case when the transition kernel of the Markov chain is state-dependent. Such stochastic optimization problems arise in various machine learning problems including strategic classification and reinforcement learning. For this problem, we study both projection-based and projection-free algorithms. In both cases, we establish that the number of calls to the stochastic first-order oracle to obtain an appropriately defined \$\epsilon\$-stationary point is of the order \$\mathcal{O}(1/\epsilon^{2.5})\$. In the projection-free setting we additionally establish that the number of calls to the linear minimization oracle is of order \$\mathcal{O}(1/\epsilon^{5.5})\$. We also empirically demonstrate the performance of our algorithm on the problem of strategic classification with neural networks.

## [Learning Probabilistic Models from Generator Latent Spaces with Hat EBM](#)

- Mitch Hill · Erik Nijkamp · Bo Pang · Jonathan Mitchell · Song-Chun Zhu
- abstract@[open-review](#): This work proposes a method for using any generator network as the foundation of an Energy-Based Model (EBM). Our formulation posits that observed images are the sum of unobserved latent variables passed through the generator network and a residual random variable that spans the gap between the generator output and the image manifold. One can then define a \emph{hat EBM} that includes the generator as part of its forward pass. The model can be trained without inferring the latent variables of the observed data or dealing with the generator Jacobian determinant. This enables explicit probabilistic modeling of the output distribution of any type of generator network. Experiments show strong performance of the proposed method on (1) unconditional ImageNet synthesis at 128\$\times\$128 resolution, (2) refining the output of existing generators, and (3) learning EBMs that incorporate non-probabilistic generators.

## [S3GC: Scalable Self-Supervised Graph Clustering](#)

- Fnu Devvrit · Aditya Sinha · Inderjit Dhillon · Prateek Jain
- abstract@[open-review](#): We study the problem of clustering graphs with additional side-information of node features. The problem is extensively studied, and several existing methods exploit Graph Neural Networks to learn node representations. However, most of the existing methods focus on generic representations instead of their cluster-ability or do not scale to large scale graph datasets. In this work, we propose S3GC which uses contrastive learning along with Graph Neural Networks and node features to learn clusterable features. We empirically demonstrate that S3GC is able to learn the correct cluster structure even when graph information or node features are individually not informative enough to learn correct clusters. Finally, using extensive evaluation on a variety of benchmarks, we demonstrate that S3GC is able to significantly outperform state-of-the-art methods in terms of clustering accuracy -- with as much as 5% gain in NMI -- while being scalable to graphs of size 100M.

## [DP-PCA: Statistically Optimal and Differentially Private PCA](#)

- Xiyang Liu · Weihao Kong · Prateek Jain · Sewoong Oh
- abstract@[open-review](#): We study the canonical statistical task of computing the principal component from i.i.d.~data under differential privacy. Although extensively studied in literature, existing solutions fall short on two key aspects: (\$i\$) even for Gaussian data, existing private algorithms require the number of samples \$n\$ to scale super-linearly with \$d\$, i.e., \$n=\Omega(d^{3/2})\$, to obtain non-trivial results while non-private PCA requires only \$n=O(d)\$, and (\$ii\$) existing techniques suffer from a large error even when the variance in each data point is small. We propose DP-PCA method that uses a single-pass minibatch gradient descent style algorithm to overcome the above limitations. For sub-Gaussian data, we provide nearly optimal statistical error rates even for \$n=O(d \log d)\$.

## [Label-invariant Augmentation for Semi-Supervised Graph Classification](#)

- Han Yue · Chunhui Zhang · Chuxu Zhang · Hongfu Liu
- abstract@[open-review](#): Recently, contrastiveness-based augmentation surges a new climax in the computer vision domain, where some operations, including rotation, crop, flip, combined with dedicated algorithms, dramatically increase the model generalization and robustness. Following this trend, some pioneering attempts employ the similar idea to graph data. Nevertheless, unlike images, it is much more difficult to design reasonable augmentations without changing the nature of graphs. Although exciting, the current graph contrastive learning does not achieve as promising performance as visual contrastive learning. We conjecture the inferior performance of graph contrastive learning might result from the violation of the label-invariant augmentation assumption. In light of this, we propose a label-invariant augmentation for graph-structured data to address this challenge. Different from the node/edge modification and subgraph extraction, we conduct the augmentation in the representation space and generate the augmented samples in the most difficult direction while keeping the label of augmented data the same as the original samples. In the semi-supervised scenario, we demonstrate our proposed method outperforms the classical graph neural network based methods and recent graph contrastive learning on eight benchmark graph-structured data, followed by several in-depth experiments to further explore the label-invariant augmentation in several aspects.

## [Implications of Model Indeterminacy for Explanations of Automated Decisions](#)

- Marc-Etienne Brunet · Ashton Anderson · Richard Zemel
- abstract@[open-review](#): There has been a significant research effort focused on explaining predictive models, for example through post-hoc explainability and recourse methods. Most of the proposed techniques operate upon a single, fixed, predictive model. However, it is well-known that given a dataset and a predictive task, there may be a multiplicity of models that solve the problem (nearly) equally well. In this work, we investigate the implications of this kind of model indeterminacy on the post-hoc explanations of predictive models. We show how it can lead to explanatory multiplicity, and we explore the underlying drivers. We show how predictive multiplicity, and the related concept of epistemic uncertainty, are not indicative of explanatory multiplicity. We further illustrate how a set of models showing very similar aggregate performance on a test dataset may show large variations in their local explanations, i.e., for a specific input. We explore these effects for Shapley value based explanations on three risk assessment datasets. Our results indicate that model indeterminacy may have a substantial impact on explanations in practice, leading to inconsistent and even contradicting explanations.

## [Associating Objects and Their Effects in Video through Coordination Games](#)

- Erika Lu · Forrester Cole · Weidi Xie · Tali Dekel · Bill Freeman · Andrew Zisserman · Michael Rubinstein
- abstract@[open-review](#): We explore a feed-forward approach for decomposing a video into layers, where each layer contains an object of interest along with its associated shadows, reflections, and other visual effects. This problem is challenging since associated effects vary widely with the 3D geometry and lighting conditions in the scene, and ground-truth labels for visual effects are difficult (and in some cases impractical) to collect. We take a self-supervised approach and train a neural network to produce a foreground image and alpha matte from a rough object segmentation mask under a reconstruction and sparsity loss. Under reconstruction loss, the layer decomposition problem is underdetermined: many combinations of layers may reconstruct the input video. Inspired by the game theory concept of focal points---or Schelling points---we pose the problem as a coordination game, where each player (network) predicts the effects for a single object without knowledge of the other players' choices. The players learn to converge on the ``natural'' layer decomposition in order to maximize the likelihood of their choices aligning with the other players'. We train the network to play this game with itself, and show how to design the rules of this game so that the focal point lies at the correct layer decomposition. We demonstrate feed-forward results on a challenging synthetic dataset, then show that pretraining on this dataset significantly reduces optimization time for real videos.

## [GraphQNTK: Quantum Neural Tangent Kernel for Graph Data](#)

- Yehui Tang · Junchi Yan
- abstract@[open-review](#): Graph Neural Networks (GNNs) and Graph Kernels (GKs) are two fundamental tools used to analyze graph-structured data. Efforts have been recently made in developing a composite graph learning architecture combining the expressive power of GNNs and the transparent trainability of GKs. However, learning efficiency on these models should be carefully considered as the huge computation overhead. Besides, their convolutional methods are often straightforward and introduce severe loss of graph structure information. In this paper, we design a novel quantum graph learning model to characterize the structural information while using quantum parallelism to improve computing efficiency. Specifically, a quantum algorithm is proposed to approximately estimate the neural tangent kernel of the underlying graph neural network where a multi-head quantum attention mechanism is introduced to properly incorporate semantic similarity information of nodes into the model. We empirically show that our method achieves competitive performance on several graph classification benchmarks, and theoretical analysis is provided to demonstrate the superiority of our quantum algorithm.

## [Near-Optimal Goal-Oriented Reinforcement Learning in Non-Stationary Environments](#)

- Liyu Chen · Haipeng Luo
- abstract@[open-review](#): We initiate the study of dynamic regret minimization for goal-oriented reinforcement learning modeled by a non-stationary stochastic shortest path problem with changing cost and transition functions. We start by establishing a lower bound  $\Omega((B_{\star} \cdot SAT_{\star}) (\Delta_c + B_{\star}^2 \Delta_P)^{1/3} K^{2/3})$ , where  $B_{\star}$  is the maximum expected cost of the optimal policy of any episode starting from any state,  $T_{\star}$  is the maximum hitting time of the optimal policy of any episode starting from the initial state,  $SA$  is the number of state-action

pairs,  $\Delta_c$  and  $\Delta_P$  are the amount of changes of the cost and transition functions respectively, and  $K$  is the number of episodes. The different roles of  $\Delta_c$  and  $\Delta_P$  in this lower bound inspire us to design algorithms that estimate costs and transitions separately. Specifically, assuming the knowledge of  $\Delta_c$  and  $\Delta_P$ , we develop a simple but sub-optimal algorithm and another more involved minimax optimal algorithm (up to logarithmic terms). These algorithms combine the ideas of finite-horizon approximation [Chen et al., 2021b], special Bernstein-style bonuses of the MVP algorithm [Zhang et al., 2020], adaptive confidence widening [Wei and Luo, 2021], as well as some new techniques such as properly penalizing long-horizon policies. Finally, when  $\Delta_c$  and  $\Delta_P$  are unknown, we develop a variant of the MASTER algorithm [Wei and Luo, 2021] and integrate the aforementioned ideas into it to achieve  $\widetilde{O}(\min\{B_{\star}\} \sqrt{ALK}, (B_{\star})^2 S^2 AT_{\star}) (\Delta_c + B_{\star} \Delta_P)^{1/3} K^{2/3})$  regret, where  $L$  is the unknown number of changes of the environment.

## [Learning Object Parts from Multiple Views for Low-shot Category Generalization](#)

- Stefan Stojanov · Anh Thai · Zixuan Huang · James Rehg
- abstract@[open-review](#): A hallmark of the deep learning era for computer vision is using large and annotated datasets training to feature representations for tasks ranging from object recognition and semantic segmentation to optical flow estimation and novel view synthesis of scenes. In this work, we aim to learn discriminative object part representations for low-shot category recognition without requiring any category labels. To this end, we propose Deep Object Part Encodings (DOPE), which can be trained from multiple views of object instances without any category or semantic object part labels. To train DOPE, we assume access to sparse depths, foreground masks and known cameras to obtain pixel-level correspondences between views of an object, and use this to formulate a self-supervised learning task to learn object parts. We find that DOPE can directly be used for low-shot classification of novel categories using local-part matching, and is competitive with and outperforms supervised and self-supervised learning baselines.

## [Human-AI Shared Control via Policy Dissection](#)

- Quanyi Li · Zhenghao Peng · Haibin Wu · Lan Feng · Bolei Zhou
- abstract@[open-review](#): Human-AI shared control allows human to interact and collaborate with autonomous agents to accomplish control tasks in complex environments. Previous Reinforcement Learning (RL) methods attempted goal-conditioned designs to achieve human-controllable policies at the cost of redesigning the reward function and training paradigm. Inspired by the neuroscience approach to investigate the motor cortex in primates, we develop a simple yet effective frequency-based approach called Policy Dissection to align the intermediate representation of the learned neural controller with the kinematic attributes of the agent behavior. Without modifying the neural controller or retraining the model, the proposed approach can convert a given RL-trained policy into a human-controllable policy. We evaluate the proposed approach on many RL tasks such as autonomous driving and locomotion. The experiments show that human-AI shared control system achieved by Policy Dissection in driving task can substantially improve the performance and safety in unseen traffic scenes. With human in the inference loop, the locomotion robots also exhibit versatile controllable motion skills even though they are only trained to move forward. Our results suggest the promising direction of implementing human-AI shared autonomy through interpreting the learned representation of the autonomous agents. Code and demo videos are available at <https://metadrive.github.io/policydissect>

## [Benefits of Permutation-Equivariance in Auction Mechanisms](#)

- Tian Qin · Fengxiang He · Dingfeng Shi · Wenbing Huang · Dacheng Tao
- abstract@[open-review](#): Designing an incentive-compatible auction mechanism that maximizes the auctioneer's revenue while minimizes the bidders' ex-post regret is an important yet intricate problem in economics. Remarkable progresses have been achieved through learning the optimal auction mechanism by neural networks. In this paper, we consider the popular {it additive valuation} and {it symmetric valuation} setting; {it i.e.}, the valuation for a set of items is defined as the sum of all items' valuations in the set, and the valuation distribution is invariant when the bidders and/or the items are permuted. We prove that permutation-equivariant neural networks have significant advantages: the permutation-equivariance decreases the expected ex-post regret, improves the model generalizability, while maintains the expected revenue invariant. This implies that the permutation-equivariance helps approach the theoretically optimal {it dominant strategy incentive compatible} condition, and reduces the required sample complexity for desired generalization. Extensive experiments fully support our theory. To our best knowledge, this is the first work towards understanding the benefits of permutation-equivariance in auction mechanisms. Code will be released publicly.

## [Mean Estimation with User-level Privacy under Data Heterogeneity](#)

- Rachel Cummings · Vitaly Feldman · Audra McMillan · Kunal Talwar
- abstract@[open-review](#): A key challenge in many modern data analysis tasks is that user data is heterogeneous. Different users may possess vastly different numbers of data points. More importantly, it cannot be assumed that all users sample from the same underlying distribution. This is true, for example in language data, where different speech styles result in data heterogeneity. In this work we propose a simple model of heterogeneous user data that differs in both distribution and quantity of data, and we provide a method for estimating the population-level mean while preserving user-level differential privacy. We demonstrate asymptotic optimality of our estimator and also prove general lower bounds on the error achievable in our problem. In particular, while the optimal non-private estimator can be shown to be linear, we show that privacy constrains us to use a non-linear estimator.

## [Contrastive Adapters for Foundation Model Group Robustness](#)

- Michael Zhang · Christopher RÃ©
- abstract@[open-review](#): While large pretrained foundation models (FMs) have shown remarkable zero-shot classification robustness to dataset-level distribution shifts, their robustness to subpopulation or group shifts is relatively underexplored. We study this problem, and find that foundation models such as CLIP may not be robust to various group shifts. Across 9 robustness benchmarks, zero-shot classification with their embeddings results in gaps of up to 80.7 percentage points (pp) between average and worst-group accuracy. Unfortunately, existing methods to improve robustness require retraining, which can be prohibitively expensive on large foundation models. We also find that efficient ways to improve model inference (e.g. via adapters, lightweight networks that transform FM embeddings) do not consistently improve and can sometimes hurt group robustness compared to zero-shot. We therefore develop an adapter training strategy to effectively and efficiently improve FM group robustness. Our motivating observation is that while poor robustness results from groups in the same class being embedded far apart in the foundation model "embedding space," standard adapter training may not actually bring these points closer together. We thus propose contrastive adapting, which contrastively trains adapters to bring sample embeddings close to both their ground-truth class embeddings and same-class sample embeddings. Across the 9 robustness benchmarks, contrastive adapting consistently improves group robustness, raising worst-group accuracy by 8.5 to 56.0 pp over zero-shot. Our approach is also efficient, doing so without any FM finetuning and only a fixed set of FM embeddings. On popular benchmarks such as Waterbirds and CelebA, this leads to worst-group accuracy comparable to state-of-the-art methods, while only training <1% of the model parameters.

## [Learning Gradient Fields for Object Arrangement](#)

- Mingdong Wu · Fangwei Zhong · Yulong Xia · Hao Dong
- abstract@[open-review](#): Object Arrangement is to move objects from shuffled layouts to a normative target distribution, e.g., tidy rooms. However, it remains challenging for AI agents, as it is hard to describe the target distribution (goal state) for reward engineering or collect expert trajectories as demonstrations. Hence, it is infeasible to directly employ reinforcement learning or imitation learning algorithms to address the task. This paper aims to search for a policy only with a set of examples from a target distribution instead of a handcrafted reward function. We employ the score-matching objective to train a target gradient field (TarGF), indicating a direction on each object to increase the likelihood of the target distribution. For object

arrangement, the TarGF can be used in two ways: 1) For model-based planning, we can cast the target gradient into a reference control, and output actions with a distributed path planner; 2) For model-free reinforcement learning, the TarGF is not only used for estimating the delta likelihood as a reward but also provides suggested actions in residual policy learning. Experimental results in ball arrangement and room arrangement demonstrate that our method significantly outperforms the state-of-the-art methods in the quality of the terminal state, the efficiency of the control process, and scalability.

## [Infinite-Fidelity Coregionalization for Physical Simulation](#)

- Shibo Li · Zheng Wang · Robert Kirby · Shandian Zhe
- abstract@[open-review](#): Multi-fidelity modeling and learning is important in physical simulation related applications. It can leverage both low-fidelity and high-fidelity examples for training so as to reduce the cost of data generation while still achieving good performance. While existing approaches only model finite, discrete fidelities, in practice, the fidelity choice is often continuous and infinite, which can correspond to a continuous mesh spacing or finite element length. In this paper, we propose Infinite Fidelity Coregionalization (IFC). Given the data, our method can extract and exploit rich information within continuous, infinite fidelities to bolster the prediction accuracy. Our model can interpolate and/or extrapolate the predictions to novel fidelities, which can be even higher than the fidelities of training data. Specifically, we introduce a low-dimensional latent output as a continuous function of the fidelity and input, and multiple it with a basis matrix to predict high-dimensional solution outputs. We model the latent output as a neural Ordinary Differential Equation (ODE) to capture the complex relationships within and integrate information throughout the continuous fidelities. We then use Gaussian processes or another ODE to estimate the fidelity-varying bases. For efficient inference, we reorganize the bases as a tensor, and use a tensor-Gaussian variational posterior to develop a scalable inference algorithm for massive outputs. We show the advantage of our method in several benchmark tasks in computational physics.

## [Automatic Differentiation of Programs with Discrete Randomness](#)

- Gaurav Arya · Moritz Schauer · Frank Schäfer · Christopher Rackauckas
- abstract@[open-review](#): Automatic differentiation (AD), a technique for constructing new programs which compute the derivative of an original program, has become ubiquitous throughout scientific computing and deep learning due to the improved performance afforded by gradient-based optimization. However, AD systems have been restricted to the subset of programs that have a continuous dependence on parameters. Programs that have discrete stochastic behaviors governed by distribution parameters, such as flipping a coin with probability \$p\$ of being heads, pose a challenge to these systems because the connection between the result (heads vs tails) and the parameters (\$p\$) is fundamentally discrete. In this manuscript we develop a new reparameterization methodology that allows for generating programs whose expectation is the derivative of the expectation of the original program. We showcase how this method gives an unbiased and low variance estimator which is as automated as traditional AD mechanisms. We demonstrate unbiased forward-mode AD of discrete-time Markov chains, agent-based models such as Conway's Game of Life, and unbiased reverse-mode AD of a particle filter.

## [Open-Ended Reinforcement Learning with Neural Reward Functions](#)

- Robert Meier · Asier Mujika
- abstract@[open-review](#): Inspired by the great success of unsupervised learning in Computer Vision and Natural Language Processing, the Reinforcement Learning community has recently started to focus more on unsupervised discovery of skills. Most current approaches, like DIAYN or DADS, optimize some form of mutual information objective. We propose a different approach that uses reward functions encoded by neural networks. These are trained iteratively to reward more complex behaviour. In high-dimensional robotic environments our approach learns a wide range of interesting skills including front-flips for Half-Cheetah and one-legged running for Humanoid. It is the first skill discovery algorithm that can learn such skills without relying on any form of feature engineering. In the pixel-based Montezuma's Revenge environment our method also works with minimal changes and it learns complex skills that involve interacting with items and visiting diverse locations.

## [MABSplit: Faster Forest Training Using Multi-Armed Bandits](#)

- Mo Tiwari · Ryan Kang · Jaeyong Lee · Chris Piech · Ilan Shomorony · Sebastian Thrun · Martin Zhang
- abstract@[open-review](#): Ensemble learning methods such as random forest are some of the most widely used machine learning models, especially in domains that necessitate interpretability. We present an algorithm that accelerates the training of random forest and other popular tree-based learning methods. At the core of our algorithm is a novel and fast node-splitting subroutine, dubbed MABSplit, used to efficiently find split points when constructing decision trees. Our algorithm borrows techniques from the multi-armed bandit literature to judiciously determine how to allocate computational power across potential split points. We provide theoretical guarantees that MABSplit improves the computational complexity from linear to logarithmic in the number of data points. Even on small datasets such as MNIST, our algorithm leads to 7x faster training (an 85% reduction in training time) without any decrease in test accuracy. We demonstrate similar speedups when the MABSplit subroutine is used across a variety of forest-based variants, such as Extremely Random Forests and Random Patches. We also show our algorithm can be used in both classification and regression tasks. Finally, MABSplit outperforms existing methods in test performance and feature importance calculations under a fixed computational budget.

## [Dynamic Sparse Network for Time Series Classification: Learning What to See](#)

- Qiao Xiao · Boqian Wu · Yu Zhang · Shiwei Liu · Mykola Pechenizkiy · Elena Mocanu · Decebal Constantin Mocanu
- abstract@[open-review](#): The receptive field (RF), which determines what hidden signals can be seen in a time series model, is critical to improve the performance for time series classification (TSC). However, the variation of signal scales across and within time series data, makes it challenging to decide proper RF sizes for TSC. In this paper, we propose a dynamic sparse network (DSN) with sparse connections for TSC, which can learn to cover various RF without cumbersome hyper-parameters tuning. The kernels in each sparse layer are sparse and can be explored under the constraint regions by dynamic sparse training, which makes it possible to reduce the resource cost. The experiment results show that the proposed DSN model can achieve the state-of-art performance on both univariate and multivariate TSC datasets with less than 50% computational cost compared with recently baseline methods, opening the path towards more accurate resource-aware methods for time series analyses. Our code is provided in the supplementary material, and it will be made available online.

## [Antigen-Specific Antibody Design and Optimization with Diffusion-Based Generative Models](#)

- Shitong Luo · Yufeng Su · Xingang Peng · Sheng Wang · Jian Peng · Jianzhu Ma
- abstract@[open-review](#): Antibodies are immune system proteins that protect the host by binding to specific antigens such as viruses and bacteria. The binding between antibodies and antigens are mainly determined by the complementarity-determining regions (CDR) on the antibodies. In this work, we develop a deep generative model that jointly models sequences and structures of CDRs based on diffusion processes and equivariant neural networks. Our method is the first deep learning-based method that can explicitly target specific antigen structures and generate antibodies at atomic resolution. The model is a "Swiss Army Knife" which is capable of sequence-structure co-design, sequence design for given backbone structures, and antibody optimization. For antibody optimization, we propose a special sampling scheme that first perturbs the given antibody and then denoises it. As the number of available antibody structures is relatively scarce, we curate a new dataset that contains antibody-like proteins as a complement to the original antibody dataset for training. We conduct extensive experiments to evaluate the quality of both sequences and structures of designed antibodies. We find that our model could yield highly competitive results in terms of binding affinity measured by biophysical energy functions and other protein design metrics.

## Cross Aggregation Transformer for Image Restoration

- Zheng Chen · Yulun Zhang · Jinjin Gu · yongbing zhang · Linghe Kong · Xin Yuan
- abstract@[open-review](#): Recently, Transformer architecture has been introduced into image restoration to replace convolution neural network (CNN) with surprising results. Considering the high computational complexity of Transformer with global attention, some methods use the local square window to limit the scope of self-attention. However, these methods lack direct interaction among different windows, which limits the establishment of long-range dependencies. To address the above issue, we propose a new image restoration model, Cross Aggregation Transformer (CAT). The core of our CAT is the Rectangle-Window Self-Attention (Rwin-SA), which utilizes horizontal and vertical rectangle window attention in different heads parallelly to expand the attention area and aggregate the features cross different windows. We also introduce the Axial-Shift operation for different window interactions. Furthermore, we propose the Locality Complementary Module to complement the self-attention mechanism, which incorporates the inductive bias of CNN (e.g., translation invariance and locality) into Transformer, enabling global-local coupling. Extensive experiments demonstrate that our CAT outperforms recent state-of-the-art methods on several image restoration applications. The code and models are available at <https://github.com/zhengchen1999/CAT>.

## Meta-Complementing the Semantics of Short Texts in Neural Topic Models

- Ce Zhang · Hady Lauw
- abstract@[open-review](#): Topic models infer latent topic distributions based on observed word co-occurrences in a text corpus. While typically a corpus contains documents of variable lengths, most previous topic models treat documents of different lengths uniformly, assuming that each document is sufficiently informative. However, shorter documents may have only a few word co-occurrences, resulting in inferior topic quality. Some other previous works assume that all documents are short, and leverage external auxiliary data, e.g., pretrained word embeddings and document connectivity. Orthogonal to existing works, we remedy this problem within the corpus itself by proposing a Meta-Complement Topic Model, which improves topic quality of short texts by transferring the semantic knowledge learned on long documents to complement semantically limited short texts. As a self-contained module, our framework is agnostic to auxiliary data and can be further improved by flexibly integrating them into our framework. Specifically, when incorporating document connectivity, we further extend our framework to complement documents with limited edges. Experiments demonstrate the advantage of our framework.

## TANKBind: Trigonometry-Aware Neural NetworKs for Drug-Protein Binding Structure Prediction

- Wei Lu · Qifeng Wu · Jixian Zhang · Jiahua Rao · Chengtao Li · Shuangjia Zheng
- abstract@[open-review](#): Illuminating interactions between proteins and small drug molecules is a long-standing challenge in the field of drug discovery. Despite the importance of understanding these interactions, most previous works are limited by hand-designed scoring functions and insufficient conformation sampling. The recently-proposed graph neural network-based methods provides alternatives to predict protein-ligand complex conformation in a one-shot manner. However, these methods neglect the geometric constraints of the complex structure and weaken the role of local functional regions. As a result, they might produce unreasonable conformations for challenging targets and generalize poorly to novel proteins. In this paper, we propose Trigonometry-Aware Neural networKs for binding structure prediction, TANKBind, that builds trigonometry constraint as a vigorous inductive bias into the model and explicitly attends to all possible binding sites for each protein by segmenting the whole protein into functional blocks. We construct novel contrastive losses with local region negative sampling to jointly optimize the binding interaction and affinity. Extensive experiments show substantial performance gains in comparison to state-of-the-art physics-based and deep learning-based methods on commonly-used benchmark datasets for both binding structure and affinity predictions with variant settings.

## Provably Efficient Model-Free Constrained RL with Linear Function Approximation

- Arnob Ghosh · Xingyu Zhou · Ness Shroff
- abstract@[open-review](#): We study the constrained reinforcement learning problem, in which an agent aims to maximize the expected cumulative reward subject to a constraint on the expected total value of a utility function. In contrast to existing model-based approaches or model-free methods accompanied with a `simulator™, we aim to develop the first \emph{model-free}, \emph{simulator-free} algorithm that achieves a sublinear regret and a sublinear constraint violation even in \emph{large-scale} systems. To this end, we consider the episodic constrained Markov decision processes with linear function approximation, where the transition dynamics and the reward function can be represented as a linear function of some known feature mapping. We show that  $\tilde{\mathcal{O}}(\sqrt{d^3H^3T})$  regret and  $\tilde{\mathcal{O}}(\sqrt{d^3H^3T})$  constraint violation bounds can be achieved, where  $d$  is the dimension of the feature mapping,  $H$  is the length of the episode, and  $T$  is the total number of steps. Our bounds are attained without explicitly estimating the unknown transition model or requiring a simulator, and they depend on the state space only through the dimension of the feature mapping. Hence our bounds hold even when the number of states goes to infinity. Our main results are achieved via novel adaptations of the standard LSVI-UCB algorithms. In particular, we first introduce primal-dual optimization into the LSVI-UCB algorithm to balance between regret and constraint violation. More importantly, we replace the standard greedy selection with respect to the state-action function with a soft-max policy. This turns out to be key in establishing uniform concentration (a critical step for provably efficient model-free exploration) for the constrained case via its approximation-smoothness trade-off. Finally, we also show that one can achieve an even zero constraint violation for large enough  $T$  by trading the regret a little bit but still maintaining the same order with respect to  $T$ .

## Exploration via Planning for Information about the Optimal Trajectory

- Viraj Mehta · Ian Char · Joseph Abbate · Rory Conlin · Mark Boyer · Stefano Ermon · Jeff Schneider · Willie Neiswanger
- abstract@[open-review](#): Many potential applications of reinforcement learning (RL) are stymied by the large numbers of samples required to learn an effective policy. This is especially true when applying RL to real-world control tasks, e.g. in the sciences or robotics, where executing a policy in the environment is costly. In popular RL algorithms, agents typically explore either by adding stochasticity to a reward-maximizing policy or by attempting to gather maximal information about environment dynamics without taking the given task into account. In this work, we develop a method that allows us to plan for exploration while taking both the task and the current knowledge about the dynamics into account. The key insight to our approach is to plan an action sequence that maximizes the expected information gain about the optimal trajectory for the task at hand. We demonstrate that our method learns strong policies with 2x fewer samples than strong exploration baselines and 200x fewer samples than model free methods on a diverse set of low-to-medium dimensional control tasks in both the open-loop and closed-loop control settings.

## Robust Bayesian Regression via Hard Thresholding

- Fan zheyi · Qingpei Hu · Zhaohui Li
- abstract@[open-review](#): By combining robust regression and prior information, we develop an effective robust regression method that can resist adaptive adversarial attacks. Due to the widespread existence of noise and data corruption, it is necessary to recover the true regression parameters when a certain proportion of the response variables have been corrupted. Methods to overcome this problem often involve robust least-squares regression. However, few methods achieve good performance when dealing with severe adaptive adversarial attacks. Based on the combination of prior information and robust regression via hard thresholding, this paper proposes an algorithm that improves the breakdown point when facing adaptive adversarial attacks. Furthermore, to improve the robustness and reduce the estimation error caused by the inclusion of a prior, the idea of Bayesian reweighting is used to construct a more robust algorithm. We prove the theoretical convergence of proposed algorithms under mild conditions. Extensive experiments show that,

under different dataset attacks, our algorithms achieve state-of-the-art results compared with other benchmark algorithms, demonstrating the robustness of the proposed approach.

## [Population geometry enables fast sampling in spiking neural networks](#)

- Paul Masset · Jacob Zavatone-Veth · J. Patrick Connor · Venkatesh Murthy · Cengiz Pehlevan
- abstract@[open-review](#): For animals to navigate an uncertain world, their brains need to estimate uncertainty at the timescales of sensations and actions. Sampling-based algorithms afford a theoretically-grounded framework for probabilistic inference in neural circuits, but it remains unknown how one can implement fast sampling algorithms in biologically-plausible spiking networks. Here, we propose to leverage the population geometry, controlled by the neural code and the neural dynamics, to implement fast samplers in spiking neural networks. We first show that two classes of spiking samplers---efficient balanced spiking networks that simulate Langevin sampling, and networks with probabilistic spike rules that implement Metropolis-Hastings sampling---can be unified within a common framework. We then show that careful choice of population geometry, corresponding to the natural space of parameters enables rapid inference of parameters drawn from strongly-correlated high-dimensional distributions in both networks. Our results suggest design principles for algorithms for sampling-based probabilistic inference in spiking neural networks, yielding potential inspiration for neuromorphic computing and testable predictions for neurobiology.

## [Understanding Non-linearity in Graph Neural Networks from the Bayesian-Inference Perspective](#)

- Rongzhe Wei · Haoteng YIN · Junteng Jia · Austin Benson · Pan Li
- abstract@[open-review](#): Graph neural networks (GNNs) have shown superiority in many prediction tasks over graphs due to their impressive capability of capturing nonlinear relations in graph-structured data. However, for node classification tasks, often, only marginal improvement of GNNs has been observed in practice over their linear counterparts. Previous works provide very few understandings of this phenomenon. In this work, we resort to Bayesian learning to give an in-depth investigation of the functions of non-linearity in GNNs for node classification tasks. Given a graph generated from the statistical model CSBM, we observe that the max-a-posterior estimation of a node label given its own and neighbors' attributes consists of two types of non-linearity, the transformation of node attributes and a ReLU-activated feature aggregation from neighbors. The latter surprisingly matches the type of non-linearity used in many GNN models. By further imposing Gaussian assumption on node attributes, we prove that the superiority of those ReLU activations is only significant when the node attributes are far more informative than the graph structure, which nicely explains previous empirical observations. A similar argument is derived when there is a distribution shift of node attributes between the training and testing datasets. Finally, we verify our theory on both synthetic and real-world networks.

## [Rethinking Alignment in Video Super-Resolution Transformers](#)

- Shuwei Shi · Jinjin Gu · Liangbin Xie · Xintao Wang · Yujiu Yang · Chao Dong
- abstract@[open-review](#): The alignment of adjacent frames is considered to be an essential operation in video super-resolution (VSR). Advanced VSR models, including the latest VSR Transformers, are generally equipped with well-designed alignment modules. However, the progress of the self-attention mechanism may violate this common sense. In this paper, we rethink the role of alignment in VSR Transformers and make several counter-intuitive observations. Our experiments show that: (i) VSR Transformers can directly utilize multi-frame information from unaligned videos, and (ii) existing alignment methods are sometimes harmful to VSR Transformers. These observations indicate that we can further improve the performance of VSR Transformers simply by removing the alignment module and adopting a larger attention window. Nevertheless, such designs will dramatically increase the computational burden, and cannot deal with large motions. Therefore, we propose a new and efficient alignment method called patch alignment, which aligns image patches instead of pixels. VSR Transformers equipped with patch alignment could demonstrate state-of-the-art performance on multiple benchmarks. Our work provides useful insights on how multi-frame information is used in VSR and how to select alignment methods for different networks/datasets.

## [GenerSpeech: Towards Style Transfer for Generalizable Out-Of-Domain Text-to-Speech](#)

- Rongjie Huang · Yi Ren · Jinglin Liu · Chenye Cui · Zhou Zhao
- abstract@[open-review](#): Style transfer for out-of-domain (OOD) speech synthesis aims to generate speech samples with unseen style (e.g., speaker identity, emotion, and prosody) derived from an acoustic reference, while facing the following challenges: 1) The highly dynamic style features in expressive voice are difficult to model and transfer; and 2) the TTS models should be robust enough to handle diverse OOD conditions that differ from the source data. This paper proposes GenerSpeech, a text-to-speech model towards high-fidelity zero-shot style transfer of OOD custom voice. GenerSpeech decomposes the speech variation into the style-agnostic and style-specific parts by introducing two components: 1) a multi-level style adaptor to efficiently model a large range of style conditions, including global speaker and emotion characteristics, and the local (utterance, phoneme, and word-level) fine-grained prosodic representations; and 2) a generalizable content adaptor with Mix-Style Layer Normalization to eliminate style information in the linguistic content representation and thus improve model generalization. Our evaluations on zero-shot style transfer demonstrate that GenerSpeech surpasses the state-of-the-art models in terms of audio quality and style similarity. The extension studies to adaptive style transfer further show that GenerSpeech performs robustly in the few-shot data setting. Audio samples are available at \url{https://GenerSpeech.github.io/}.

## [Leveraging Inter-Layer Dependency for Post -Training Quantization](#)

- changbao wang · DanDan Zheng · Yuanliu Liu · Liang Li
- abstract@[open-review](#): Prior works on Post-training Quantization (PTQ) typically separate a neural network into sub-nets and quantize them sequentially. This process pays little attention to the dependency across the sub-nets, hence is less optimal. In this paper, we propose a novel Network-Wise Quantization (NWQ) approach to fully leveraging inter-layer dependency. NWQ faces a larger scale combinatorial optimization problem of discrete variables than in previous works, which raises two major challenges: over-fitting and discrete optimization problem. NWQ alleviates over-fitting via a Activation Regularization (AR) technique, which better controls the activation distribution. To optimize discrete variables, NWQ introduces Annealing Softmax (ASoftmax) and Annealing Mixup (AMixup) to progressively transition quantized weights and activations from continuity to discretization, respectively. Extensive experiments demonstrate that NWQ outperforms previous state-of-the-art by a large margin: 20.24% for the challenging configuration of MobileNetV2 with 2 bits on ImageNet, pushing extremely low-bit PTQ from feasibility to usability. In addition, NWQ is able to achieve competitive results with only 10% computation cost of previous works.

## [Spectrum Random Masking for Generalization in Image-based Reinforcement Learning](#)

- Yangru Huang · Peixi Peng · Yifan Zhao · Guangyao Chen · Yonghong Tian
- abstract@[open-review](#): Generalization in image-based reinforcement learning (RL) aims to learn a robust policy that could be applied directly on unseen visual environments, which is a challenging task since agents usually tend to overfit to their training environment. To handle this problem, a natural approach is to increase the data diversity by image based augmentations. However, different with most vision tasks such as classification and detection, RL tasks are not always invariant to spatial based augmentations due to the entanglement of environment dynamics and visual appearance. In this paper, we argue with two principles for augmentations in RL: First, the augmented observations should facilitate learning a universal policy, which is robust to various distribution shifts. Second, the augmented data should be invariant to the learning signals such as action and reward. Following these rules, we revisit image-based RL tasks from the view of frequency domain and propose a novel augmentation method, namely Spectrum Random Masking (SRM), which is able to help agents to learn the whole frequency spectrum of observation for coping with various distributions and compatible with the

pre-collected action and reward corresponding to original observation. Extensive experiments conducted on DMControl Generalization Benchmark demonstrate the proposed SRM achieves the state-of-the-art performance with strong generalization potentials.

## [MetaMask: Revisiting Dimensional Confounder for Self-Supervised Learning](#)

- Jiangmeng Li · Wenwen Qiang · Yanan Zhang · Wenyi Mo · Changwen Zheng · Bing Su · Hui Xiong
- abstract@[open-review](#): As a successful approach to self-supervised learning, contrastive learning aims to learn invariant information shared among distortions of the input sample. While contrastive learning has yielded continuous advancements in sampling strategy and architecture design, it still remains two persistent defects: the interference of task-irrelevant information and sample inefficiency, which are related to the recurring existence of trivial constant solutions. From the perspective of dimensional analysis, we find out that the dimensional redundancy and dimensional confounder are the intrinsic issues behind the phenomena, and provide experimental evidence to support our viewpoint. We further propose a simple yet effective approach MetaMask, short for the dimensional Mask learned by Meta-learning, to learn representations against dimensional redundancy and confounder. MetaMask adopts the redundancy-reduction technique to tackle the dimensional redundancy issue and innovatively introduces a dimensional mask to reduce the gradient effects of specific dimensions containing the confounder, which is trained by employing a meta-learning paradigm with the objective of improving the performance of masked representations on a typical self-supervised task. We provide solid theoretical analyses to prove MetaMask can obtain tighter risk bounds for downstream classification compared to typical contrastive methods. Empirically, our method achieves state-of-the-art performance on various benchmarks.

## [Degradation-Aware Unfolding Half-Shuffle Transformer for Spectral Compressive Imaging](#)

- Yuanhao Cai · Jing Lin · Haoqian Wang · Xin Yuan · Henghui Ding · Yulun Zhang · Radu Timofte · Luc V Gool
- abstract@[open-review](#): In coded aperture snapshot spectral compressive imaging (CASSI) systems, hyperspectral image (HSI) reconstruction methods are employed to recover the spatial-spectral signal from a compressed measurement. Among these algorithms, deep unfolding methods demonstrate promising performance but suffer from two issues. Firstly, they do not estimate the degradation patterns and ill-posedness degree from the highly related CASSI to guide the iterative learning. Secondly, they are mainly CNN-based, showing limitations in capturing long-range dependencies. In this paper, we propose a principled Degradation-Aware Unfolding Framework (DAUF) that estimates parameters from the compressed image and physical mask, and then uses these parameters to control each iteration. Moreover, we customize a novel Half-Shuffle Transformer (HST) that simultaneously captures local contents and non-local dependencies. By plugging HST into DAUF, we establish the first Transformer-based deep unfolding method, Degradation-Aware Unfolding Half-Shuffle Transformer (DAUHST), for HSI reconstruction. Experiments show that DAUHST significantly surpasses state-of-the-art methods while requiring cheaper computational and memory costs. Code and models will be released to the public.

## [Compressible-composable NeRF via Rank-residual Decomposition](#)

- Jiaxiang Tang · Xiaokang Chen · Jingbo Wang · Gang Zeng
- abstract@[open-review](#): Neural Radiance Field (NeRF) has emerged as a compelling method to represent 3D objects and scenes for photo-realistic rendering. However, its implicit representation causes difficulty in manipulating the models like the explicit mesh representation. Several recent advances in NeRF manipulation are usually restricted by a shared renderer network, or suffer from large model size. To circumvent the hurdle, in this paper, we present a neural field representation that enables efficient and convenient manipulation of models. To achieve this goal, we learn a hybrid tensor rank decomposition of the scene without neural networks. Motivated by the low-rank approximation property of the SVD algorithm, we propose a rank-residual learning strategy to encourage the preservation of primary information in lower ranks. The model size can then be dynamically adjusted by rank truncation to control the levels of detail, achieving near-optimal compression without extra optimization. Furthermore, different models can be arbitrarily transformed and composed into one scene by concatenating along the rank dimension. The growth of storage cost can also be mitigated by compressing the unimportant objects in the composed scene. We demonstrate that our method is able to achieve comparable rendering quality to state-of-the-art methods, while enabling extra capability of compression and composition. Code will be made publicly available.

## [Embrace the Gap: VAEs Perform Independent Mechanism Analysis](#)

- Patrik Reizinger · Luigi Gresele · Jack Brady · Julius von Kägelgen · Dominik Zietlow · Bernhard Schälkopf · Georg Martius · Wieland Brendel · Michel Besserve
- abstract@[open-review](#): Variational autoencoders (VAEs) are a popular framework for modeling complex data distributions; they can be efficiently trained via variational inference by maximizing the evidence lower bound (ELBO), at the expense of a gap to the exact (log-)marginal likelihood. While VAEs are commonly used for representation learning, it is unclear why ELBO maximization would yield useful representations, since unregularized maximum likelihood estimation cannot invert the data-generating process. Yet, VAEs often succeed at this task. We seek to elucidate this apparent paradox by studying nonlinear VAEs in the limit of near-deterministic decoders. We first prove that, in this regime, the optimal encoder approximately inverts the decoder---a commonly used but unproven conjecture---which we refer to as self-consistency. Leveraging self-consistency, we show that the ELBO converges to a regularized log-likelihood. This allows VAEs to perform what has recently been termed independent mechanism analysis (IMA): it adds an inductive bias towards decoders with column-orthogonal Jacobians, which helps recovering the true latent factors. The gap between ELBO and log-likelihood is therefore welcome, since it bears unanticipated benefits for nonlinear representation learning. In experiments on synthetic and image data, we show that VAEs uncover the true latent factors when the data generating process satisfies the IMA assumption.

## [Accelerated Linearized Laplace Approximation for Bayesian Deep Learning](#)

- Zhijie Deng · Feng Zhou · Jun Zhu
- abstract@[open-review](#): Laplace approximation (LA) and its linearized variant (LLA) enable effortless adaptation of pretrained deep neural networks to Bayesian neural networks. The generalized Gauss-Newton (GGN) approximation is typically introduced to improve their tractability. However, LA and LLA are still confronted with non-trivial inefficiency issues and should rely on Kronecker-factored, diagonal, or even last-layer approximate GGN matrices in practical use. These approximations are likely to harm the fidelity of learning outcomes. To tackle this issue, inspired by the connections between LLA and neural target kernels (NTKs), we develop a Nyström approximation to NTKs to accelerate LLA. Our method benefits from the capability of popular deep learning libraries for forward mode automatic differentiation, and enjoys reassuring theoretical guarantees. Extensive studies reflect the merits of the proposed method in aspects of both scalability and performance. Our method can even scale up to architectures like vision transformers. We also offer valuable ablation studies to diagnose our method.

## [Decoupling Knowledge from Memorization: Retrieval-augmented Prompt Learning](#)

- Xiang Chen · Lei Li · Ningyu Zhang · Xiaozhuan Liang · Shumin Deng · Chuanqi Tan · Fei Huang · Luo Si · Huajun Chen
- abstract@[open-review](#): Prompt learning approaches have made waves in natural language processing by inducing better few-shot performance while they still follow a parametric-based learning paradigm; the oblivion and rote memorization problems in learning may encounter unstable generalization issues. Specifically, vanilla prompt learning may struggle to utilize atypical instances by rote during fully-supervised training or overfit shallow patterns with low-shot data. To alleviate such limitations, we develop RetroPrompt with the motivation of decoupling knowledge from memorization to help the model strike a balance between generalization and memorization. In contrast with vanilla prompt learning, RetroPrompt constructs an open-book knowledge-store from training instances and implements a retrieval mechanism during the process of input, training and inference, thus equipping the model with the ability to retrieve related contexts from the training corpus as cues for enhancement. Extensive experiments demonstrate that RetroPrompt can obtain

better performance in both few-shot and zero-shot settings. Besides, we further illustrate that our proposed RetroPrompt can yield better generalization abilities with new datasets. Detailed analysis of memorization indeed reveals RetroPrompt can reduce the reliance of language models on memorization; thus, improving generalization for downstream tasks.

## [Improving Out-of-distribution Robustness by Adversarial Training with Structured Priors](#)

- Qixun Wang · Yifei Wang · Hong Zhu · Yisen Wang
- abstract@[open-review](#): Deep models often fail to generalize well in test domains when the data distribution differs from that in the training domain. Among numerous approaches to address this Out-of-Distribution (OOD) generalization problem, there has been a growing surge of interest in exploiting the input-robustness obtained by Adversarial Training (AT) to improve OOD performances. Recent works have revealed that the robust model obtained by conducting sample-wise AT also retains transferability to biased test domains. In this paper, we empirically show that sample-wise AT has limited improvement on OOD performance. Specifically, we find that AT can only maintain performance at smaller scales of perturbation while Universal AT (UAT) are more robust to larger-scale perturbations. This provides us with clues that the adversarial perturbations with universal (low dimensional) structures can enhance the robustness to large data distribution shifts which are common in OOD scenarios. Inspired by this, we propose two AT variants with low-rank structures to train OOD-robust models. Extensive experiments on DomainBed benchmark show that our proposed approaches outperform Empirical Risk Minimization (ERM) and sample-wise AT.

## [Width and Depth Guidelines for Deep Q-Learning: A Function Approximation Perspective](#)

- Fanghui Liu · Luca Viano · Volkan Cevher
- abstract@[open-review](#): This paper provides a theoretical study of deep neural function approximation in reinforcement learning (RL) employed with the  $\$epsilon$ -greedy exploration under the online setting. This problem setting is motivated by the successful deep Q-networks (DQN) framework that works in this regime. We provide an initial attempt on theoretical understanding deep RL from the perspective of function class and neural networks architectures (e.g., width and depth). In this work, we focus on value based algorithm with the  $\$epsilon$ -greedy exploration under Besov and Barron function spaces, which aims at approximating an  $\alpha$ -smooth Q-function in a  $d$ -dimensional feature space with  $T$  episodes. We prove that scaling the width  $m = \widetilde{\mathcal{O}}(T^{\frac{d}{2\alpha+d}})$  and the depth  $L = \widetilde{\mathcal{O}}(1)$  of the neural network for deep RL is sufficient for learning with sublinear regret in Besov spaces. Moreover, for a two layer neural network in the Barron space, scaling the width even in sublinearly in  $T$  is sufficient. In particular, for low smooth Q-function, the theory suggests that the depth should be chosen as the logarithm of the width. This matches the practical DQN setting. Our analysis relies on transforming the estimation of temporal difference error to a generalization problem under the conditionally independent but non-identically distributed data setting over the averaged empirical measure. This might have its own interest in reinforcement learning theory for better understanding  $\$epsilon$ -greedy exploration in deep RL.

## [Unsupervised Learning for Combinatorial Optimization with Principled Objective Design](#)

- Haoyu Peter Wang · Nan Wu · Hang Yang · Cong Hao · Pan Li
- abstract@[open-review](#): Using machine learning to solve combinatorial optimization (CO) problems is challenging, especially when the data is unlabeled. This work proposes an unsupervised learning framework for CO problems. Our framework follows the standard relaxation-plus-rounding approach and adopts neural networks to parameterize the relaxed solutions so that simple back-propagation can train them end-to-end. Our key contribution is the observation that if the relaxed objective satisfies entry-wise concavity, a low optimization loss guarantees the quality of the obtained integral solutions. This observation significantly generalizes the applicability of the previous framework inspired by Erdos' probabilistic method (Karalias & Loukas, 2020). Our framework is particularly suitable to guide the design of objective models in the applications where the objectives are not given explicitly while requiring being modeled and learned first. We evaluate our framework by solving a synthetic graph optimization problem, and two real-world applications including resource allocation in circuit design and approximate computing. Our framework largely outperforms the baselines based on reinforcement learning and Gumbel-softmax tricks.

## [Adversarial Robustness is at Odds with Lazy Training](#)

- Yunjuan Wang · Enayat Ullah · Poorya Mianjy · Raman Arora
- abstract@[open-review](#): Recent works show that random neural networks are vulnerable against adversarial attacks [Daniely and Schacham, 2020] and that such attacks can be easily found using a single step of gradient descent [Bubeck et al., 2021]. In this work, we take one step further and show that one single gradient step can find adversarial examples for networks trained in the so-called lazy regime. This regime is interesting because even though the neural network weights remain close to the initialization, there exist networks with small generalization error, which can be found efficiently using first-order methods. Our work challenges the model of the lazy regime, the only regime in which neural networks are provably efficiently learnable. We show that the networks trained in this regime, even though they enjoy good theoretical computational guarantees, remain vulnerable to adversarial examples. In doing so, we resolve an open question posed by the work of [Bubeck et al., 2021], who show a similar result for random neural networks. To the best of our knowledge, this is the first work to prove that such well-generalizable neural networks are still vulnerable to adversarial attacks.

## [Differentially Private Generalized Linear Models Revisited](#)

- Raman Arora · Raef Bassily · Cristbal Guzman · Michael Menart · Enayat Ullah
- abstract@[open-review](#): We study the problem of  $(\epsilon, \delta)$ -differentially private learning of linear predictors with convex losses. We provide results for two subclasses of loss functions. The first case is when the loss is smooth and non-negative but not necessarily Lipschitz (such as the squared loss). For this case, we establish an upper bound on the excess population risk of  $\tilde{\Omega}(\frac{1}{n}\sqrt{\frac{1}{\epsilon}} + \min\{\frac{1}{\epsilon}, \frac{1}{\delta}\})$ , where  $n$  is the number of samples,  $d$  is the dimension of the problem, and  $w^*$  is the minimizer of the population risk. Apart from the dependence on  $\|w^*\|$ , our bound is essentially tight in all parameters. In particular, we show a lower bound of  $\Omega(\frac{1}{n}\sqrt{\frac{1}{\epsilon}} + \min\{\frac{1}{\epsilon}, \frac{1}{\delta}\})$ . We also revisit the previously studied case of Lipschitz losses [cite{SSTT21}](#). For this case, we close the gap in the existing work and show that the optimal rate is (up to log factors)  $\Theta(\frac{1}{n}\sqrt{\frac{1}{\epsilon}} + \min\{\frac{1}{\epsilon}, \frac{1}{\delta}\})$ , where  $\text{rank}(X)$  is the rank of the design matrix. This improves over existing work in the high privacy regime. Finally, our algorithms involve a private model selection approach that we develop to enable attaining the stated rates without *a-priori* knowledge of  $\|w^*\|$ .

## [NeuForm: Adaptive Overfitting for Neural Shape Editing](#)

- Connor Lin · Niloy Mitra · Gordon Wetzstein · Leonidas Guibas · Paul Guerrero
- abstract@[open-review](#): Neural representations are popular for representing shapes as they can be used for data cleanup, model completion, shape editing, and shape synthesis. Current neural representations can be categorized as either overfitting to a single object instance, or representing a collection of objects. However, neither allows accurate editing of neural scene representations: on the one hand, methods that overfit objects achieve highly accurate reconstructions but do not support editing, as they do not generalize to unseen object configurations; on the other hand, methods that represent a family of objects with variations do generalize but produce approximate reconstructions. We propose NeuForm to combine the advantages of both overfitted and generalizable representations by adaptively overfitting a generalizable representation to regions where reliable data is available, while using the generalizable representation everywhere else. We achieve this with a carefully designed architecture and an approach that blends the network weights of

the two representations. We demonstrate edits that successfully reconfigure parts of human-made shapes, such as chairs, tables, and lamps, while preserving the accuracy of an overfitted shape representation. We compare with two state-of-the-art competitors and demonstrate clear improvements in terms of plausibility and fidelity of the resultant edits.

## [Improving Multi-Task Generalization via Regularizing Spurious Correlation](#)

- Ziniu Hu · Zhe Zhao · Xinyang Yi · Tiansheng Yao · Lichan Hong · Yizhou Sun · Ed Chi
- abstract@[open-review](#): Multi-Task Learning (MTL) is a powerful learning paradigm to improve generalization performance via knowledge sharing. However, existing studies find that MTL could sometimes hurt generalization, especially when two tasks are less correlated. One possible reason that hurts generalization is spurious correlation, i.e., some knowledge is spurious and not causally related to task labels, but the model could mistakenly utilize them and thus fail when such correlation changes. In MTL setup, there exist several unique challenges of spurious correlation. First, the risk of having non-causal knowledge is higher, as the shared MTL model needs to encode all knowledge from different tasks, and causal knowledge for one task could be potentially spurious to the other. Second, the confounder between task labels brings in a different type of spurious correlation to MTL. Given such label-label confounders, we theoretically and empirically show that MTL is prone to taking non-causal knowledge from other tasks. To solve this problem, we propose Multi-Task Causal Representation Learning (MT-CRL) framework. MT-CRL aims to represent multi-task knowledge via disentangled neural modules, and learn which module is causally related to each task via MTL-specific invariant regularization. Experiments show that MT-CRL could enhance MTL model's performance by 5.5% on average over Multi-MNIST, MovieLens, Taskonomy, CityScape, and NYUv2, and show it could indeed alleviate spurious correlation problem.

## [Falsification before Extrapolation in Causal Effect Estimation](#)

- Michael Oberst · Zeshan M Hussain · Ming-Chieh Shih · David Sontag
- abstract@[open-review](#): Randomized Controlled Trials (RCTs) represent a gold standard when developing policy guidelines. However, RCTs are often narrow, and lack data on broader populations of interest. Causal effects in these populations are often estimated using observational datasets, which may suffer from unobserved confounding and selection bias. Given a set of observational estimates (e.g., from multiple studies), we propose a meta-algorithm that attempts to reject observational estimates that are biased. We do so using validation effects, causal effects that can be inferred from both RCT and observational data. After rejecting estimators that do not pass this test, we generate conservative confidence intervals on the extrapolated causal effects for subgroups not observed in the RCT. Under the assumption that at least one observational estimator is asymptotically normal and consistent for both the validation and extrapolated effects, we provide guarantees on the coverage probability of the intervals output by our algorithm. To facilitate hypothesis testing in settings where causal effect transportation across datasets is necessary, we give conditions under which a doubly-robust estimator of group average treatment effects is asymptotically normal, even when flexible machine learning methods are used for estimation of nuisance parameters. We illustrate the properties of our approach on semi-synthetic experiments based on the IHDP dataset, and show that it compares favorably to standard meta-analysis techniques.

## [Augmentations in Hypergraph Contrastive Learning: Fabricated and Generative](#)

- Tianxin Wei · Yuning You · Tianlong Chen · Yang Shen · Jingrui He · Zhangyang Wang
- abstract@[open-review](#): This paper targets at improving the generalizability of hypergraph neural networks in the low-label regime, through applying the contrastive learning approach from images/graphs. The question we focus on here is: How to construct contrastive views of hypergraphs via augmentations? We provide the solutions in two folds. First, guided by domain knowledge, we fabricate two schemes to augment hyperedges with higher-order relations encoded, and adopt three vertex augmentation strategies from graph-structured data. Then, in search of more effective views in a data-driven manner, we are the first to propose hypergraph generative models to generate augmented views, and then an end-to-end differentiable pipeline to jointly perform hypergraph augmentation and contrastive learning. Our technical innovations are reflected in designing both fabricated and generative augmentations of hypergraphs. The experimental findings include: (i) Among fabricated augmentations, augmenting hyperedges performs the best in most cases, implying that higher-order information in structures is usually more downstream-relevant; (ii) Generative augmentations do better in preserving higher-order information to further benefit generalizability; (iii) The proposed framework HyperGCL also boosts robustness and fairness of hypergraph representation learning. Codes will be made public upon acceptance.

## [Improved Bounds on Neural Complexity for Representing Piecewise Linear Functions](#)

- Kuan-Lin Chen · Harinath Garudadri · Bhaskar D Rao
- abstract@[open-review](#): A deep neural network using rectified linear units represents a continuous piecewise linear (CPWL) function and vice versa. Recent results in the literature estimated that the number of neurons needed to exactly represent any CPWL function grows exponentially with the number of pieces or exponentially in terms of the factorial of the number of distinct linear components. Moreover, such growth is amplified linearly with the input dimension. These existing results seem to indicate that the cost of representing a CPWL function is expensive. In this paper, we propose much tighter bounds and establish a polynomial time algorithm to find a network satisfying these bounds for any given CPWL function. We prove that the number of hidden neurons required to exactly represent any CPWL function is at most a quadratic function of the number of pieces. In contrast to all previous results, this upper bound is invariant to the input dimension. Besides the number of pieces, we also study the number of distinct linear components in CPWL functions. When such a number is also given, we prove that the quadratic complexity turns into bilinear, which implies a lower neural complexity because the number of distinct linear components is always not greater than the minimum number of pieces in a CPWL function. When the number of pieces is unknown, we prove that, in terms of the number of distinct linear components, the neural complexity of any CPWL function is at most polynomial growth for low-dimensional inputs and a factorial growth for the worst-case scenario, which are significantly better than existing results in the literature.

## [Improving Transformer with an Admixture of Attention Heads](#)

- Tan Nguyen · Tam Nguyen · Hai Do · Khai Nguyen · Vishwanath Saragadam · Minh Pham · Khuong Duy Nguyen · Nhat Ho · Stanley Osher
- abstract@[open-review](#): Transformers with multi-head self-attention have achieved remarkable success in sequence modeling and beyond. However, they suffer from high computational and memory complexities for computing the attention matrix at each head. Recently, it has been shown that those attention matrices lie on a low-dimensional manifold and, thus, are redundant. We propose the Transformer with a Finite Admixture of Shared Heads (FiSHformers), a novel class of efficient and flexible transformers that allow the sharing of attention matrices between attention heads. At the core of FiSHformer is a novel finite admixture model of shared heads (FiSH) that samples attention matrices from a set of global attention matrices. The number of global attention matrices is much smaller than the number of local attention matrices generated. FiSHformers directly learn these global attention matrices rather than the local ones as in other transformers, thus significantly improving the computational and memory efficiency of the model. We empirically verify the advantages of the FiSHformer over the baseline transformers in a wide range of practical applications including language modeling, machine translation, and image classification. On the WikiText-103, IWSLT'14 De-En and WMT'14 En-De, FiSHformers use much fewer floating-point operations per second (FLOPs), memory, and parameters compared to the baseline transformers.

## [UQGAN: A Unified Model for Uncertainty Quantification of Deep Classifiers trained via Conditional GANs](#)

- Philipp Oberdiek · Gernot Fink · Matthias Rottmann
- abstract@[open-review](#): We present an approach to quantifying both aleatoric and epistemic uncertainty for deep neural networks in image classification, based on generative adversarial networks (GANs). While most works in the literature that use GANs to generate out-of-distribution (OoD) examples only

focus on the evaluation of OoD detection, we present a GAN based approach to learn a classifier that produces proper uncertainties for OoD examples as well as for false positives (FPs). Instead of shielding the entire in-distribution data with GAN generated OoD examples which is state-of-the-art, we shield each class separately with out-of-class examples generated by a conditional GAN and complement this with a one-vs-all image classifier. In our experiments, in particular on CIFAR10, CIFAR100 and Tiny ImageNet, we improve over the OoD detection and FP detection performance of state-of-the-art GAN-training based classifiers. Furthermore, we also find that the generated GAN examples do not significantly affect the calibration error of our classifier and result in a significant gain in model accuracy.

## [A Lagrangian Duality Approach to Active Learning](#)

- Juan Elenter · Navid Naderizadeh · Alejandro Ribeiro
- abstract@[open-review](#): We consider the batch active learning problem, where only a subset of the training data is labeled, and the goal is to query a batch of unlabeled samples to be labeled so as to maximally improve model performance. We formulate the learning problem using constrained optimization, where each constraint bounds the performance of the model on labeled samples. Considering a primal-dual approach, we optimize the primal variables, corresponding to the model parameters, as well as the dual variables, corresponding to the constraints. As each dual variable indicates how significantly the perturbation of the respective constraint affects the optimal value of the objective function, we use it as a proxy of the informativeness of the corresponding training sample. Our approach, which we refer to as Active Learning via Lagrangian dualitY, or ALLY, leverages this fact to select a diverse set of unlabeled samples with the highest estimated dual variables as our query set. We demonstrate the benefits of our approach in a variety of classification and regression tasks and also discuss its limitations depending on the capacity of the model used. We also show that ALLY can be used in a generative mode to create novel maximally-informative samples.

## [Submodular Maximization in Clean Linear Time](#)

- Wenxin Li · Moran Feldman · Ehsan Kazemi · Amin Karbasi
- abstract@[open-review](#): In this paper, we provide the first deterministic algorithm that achieves  $1/2\$$ -approximation for submodular maximization subject to a knapsack constraint, while making a number of queries that scales only linearly with the size of the ground set  $\$n\$$ . Moreover, our result automatically paves the way to developing a linear-time deterministic algorithm that achieves the tight  $1-1/e\$$  approximation guarantee for submodular maximization under a cardinality (size) constraint. To complement our positive results, we also show strong information-theoretic lower bounds. More specifically, we show that when the maximum cardinality allowed for a solution is constant, no deterministic or randomized algorithm making a sub-linear number of function evaluations can guarantee any constant approximation ratio. Furthermore, when the constraint allows the selection of a constant fraction of the ground set, we show that any algorithm making fewer than  $\$O(\Omega(n/\log(n)))\$$  function evaluations cannot perform better than an algorithm that simply outputs a uniformly random subset of the ground set of the right size. Finally, we extend our results to the general case of maximizing a monotone submodular function subject to the intersection of a  $\$p\$$ -set system and multiple knapsack constraints. We extensively evaluate the performance of our algorithms on multiple real-life machine learning applications, including movie recommendation, location summarization, Twitter text summarization, and video summarization.

## [Fair Infinitesimal Jackknife: Mitigating the Influence of Biased Training Data Points Without Refitting](#)

- Prasanna Sattigeri · Soumya Ghosh · Inkit Padhi · Pierre Dognin · Kush Varshney
- abstract@[open-review](#): In consequential decision-making applications, mitigating unwanted biases in machine learning models that yield systematic disadvantage to members of groups delineated by sensitive attributes such as race and gender is one key intervention to strive for equity. Focusing on demographic parity and equality of opportunity, in this paper we propose an algorithm that improves the fairness of a pre-trained classifier by simply dropping carefully selected training data points. We select instances based on their influence on the fairness metric of interest, computed using an infinitesimal jackknife-based approach. The dropping of training points is done in principle, but in practice does not require the model to be refit. Crucially, we find that such an intervention does not substantially reduce the predictive performance of the model but drastically improves the fairness metric. Through careful experiments, we evaluate the effectiveness of the proposed approach on diverse tasks and find that it consistently improves upon existing alternatives.

## [A Simple and Optimal Policy Design for Online Learning with Safety against Heavy-tailed Risk](#)

- Feng Zhu · Zeyu Zheng · David Simchi-Levi
- abstract@[open-review](#): We consider the classical multi-armed bandit problem and design simple-to-implement new policies that simultaneously enjoy two properties: worst-case optimality for the expected regret, and safety against heavy-tailed risk for the regret distribution. Recently, Fan and Glynn (2021) showed that information-theoretic optimized bandit policies as well as general UCB policies suffer from some serious heavy-tailed risk; that is, the probability of incurring a linear regret slowly decays at a polynomial rate of  $\$1/T\$$ , as  $\$T\$$  (the time horizon) increases. Inspired by their result, we further show that any policy that incurs an instance-dependent  $\$O(\ln T)\$$  regret must incur a linear regret with probability  $\$O(\Omega(\text{poly}(1/T)))\$$  and that the heavy-tailed risk actually exists for all "instance-dependent consistent" policies. Next, for the two-armed bandit setting, we provide a simple policy design that (i) has the worst-case optimality for the expected regret at order  $\tilde{O}(\sqrt{T})$  and (ii) has the worst-case tail probability of incurring a linear regret decay at an exponential rate  $\exp(-\Omega(\sqrt{T}))$ . We further prove that this exponential decaying rate of the tail probability is optimal across all policies that have worst-case optimality for the expected regret. Finally, we generalize the policy design and analysis to the general setting with an arbitrary  $\$K\$$  number of arms. We provide detailed characterization of the tail probability bound for any regret threshold under our policy design. Numerical experiments are conducted to illustrate the theoretical findings. Our results reveal insights on the incompatibility between consistency and light-tailed risk, whereas indicate that worst-case optimality on expected regret and light-tailed risk are compatible. Our policy design and proof techniques may have the potential to be adapted to other online learning tasks that prefer to have light-tailed risk on regret distribution.

## [Robust Learning against Relational Adversaries](#)

- Yizhen Wang · Mohannad Alhanahnah · Xiaozhu Meng · Ke Wang · Mihai Christodorescu · Somesh Jha
- abstract@[open-review](#): Test-time adversarial attacks have posed serious challenges to the robustness of machine-learning models, and in many settings the adversarial perturbation need not be bounded by small  $\$l_p\$$ -norms. Motivated by attacks in program analysis and security tasks, we investigate  $\{\text{relational adversaries}\}$ , a broad class of attackers who create adversarial examples in a reflexive-transitive closure of a logical relation. We analyze the conditions for robustness against relational adversaries and investigate different levels of robustness-accuracy trade-off due to various patterns in a relation. Inspired by the insights, we propose  $\{\text{normalize-and-predict}\}$ , a learning framework that leverages input normalization to achieve provable robustness. The framework solves the pain points of adversarial training against relational adversaries and can be combined with adversarial training for the benefits of both approaches. Guided by our theoretical findings, we apply our framework to source code authorship attribution and malware detection. Results of both tasks show our learning framework significantly improves the robustness of models against relational adversaries. In the process, it outperforms adversarial training, the most noteworthy defense mechanism, by a wide margin.

## [Natural image synthesis for the retina with variational information bottleneck representation](#)

- Babak Rahmani · Demetri Psaltis · Christophe Moser
- abstract@[open-review](#): In the early visual system, high dimensional natural stimuli are encoded into the trains of neuronal spikes that transmit the information to the brain to produce perception. But, is all the visual scene information required to explain the neuronal responses? In this work, we search

for answers to this question by developing a joint model of the natural visual input and neuronal responses using the Information Bottleneck (IB) framework that is able to represent features of the input data into a few latent variables that play a role in the prediction of the outputs. The correlations between data samples, acquired from published experiments on ex-vivo retinas, is accounted for in the model by a Gaussian Process (GP) prior. The proposed IB-GP model performs competitive to the state-of-the-art feedforward convolutional networks in prediction of spike responses to natural stimuli. Finally, the IB-GP model is used in a closed loop iterative process to obtain reduced-complexity inputs that elicit responses as those elicited by the original stimuli. We found three properties of the retina's IB-GP model. First, the reconstructed stimuli from the latent variables show robustness in spike prediction across models. Second, surprisingly the dynamics of the high-dimensional stimuli and RGCs' responses are very well represented in the embeddings of the IB-GP model. Third, the minimum stimuli consist of different patterns: Gabor-type locally high-frequency filters, on- and off-center Gaussians, or a mixture of both. Overall, the IB-GP model not only provides a principled approach for joint learning of the stimuli and retina codes, which could help understand the computation of the early visual system, but could also be potentially used in the closed loop with the visual prostheses to increase their efficiency.

## [Minimax Optimal Online Imitation Learning via Replay Estimation](#)

- Gokul Swamy · Nived Rajaraman · Matt Peng · Sanjiban Choudhury · J. Bagnell · Steven Wu · Jiantao Jiao · Kannan Ramchandran
- abstract@[open-review](#): Online imitation learning is the problem of how best to mimic expert demonstrations, given access to the environment or an accurate simulator. Prior work has shown that in the \textit{infinite} sample regime, exact moment matching achieves value equivalence to the expert policy. However, in the \textit{finite} sample regime, even if one has no optimization error, empirical variance can lead to a performance gap that scales with  $H^2 / N_{\text{exp}}$  for behavioral cloning and  $H / N_{\text{exp}}$  for online moment matching, where  $H$  is the horizon and  $N_{\text{exp}}$  is the size of the expert dataset. We introduce the technique of ``replay estimation'' to reduce this empirical variance: by repeatedly executing cached expert actions in a stochastic simulator, we compute a smoother expert visitation distribution estimate to match. In the presence of general function approximation, we prove a meta theorem reducing the performance gap of our approach to the \textit{parameter estimation error} for offline classification (i.e. learning the expert policy). In the tabular setting or with linear function approximation, our meta theorem shows that the performance gap incurred by our approach achieves the optimal  $\widetilde{O}(\min(H^{3/2} / N_{\text{exp}}, H / \sqrt{N_{\text{exp}}}))$  dependency, under significantly weaker assumptions compared to prior work, Rajaraman et al. (2021). We implement multiple instantiations of our approach on several continuous control tasks and find that we are able to significantly improve policy performance across a variety of dataset sizes.

## [Tight Lower Bounds on Worst-Case Guarantees for Zero-Shot Learning with Attributes](#)

- Alessio Mazzetto · Cristina Menghini · Andrew Yuan · Eli Upfal · Stephen Bach
- abstract@[open-review](#): We develop a rigorous mathematical analysis of zero-shot learning with attributes. In this setting, the goal is to label novel classes with no training data, only detectors for attributes and a description of how those attributes are correlated with the target classes, called the class-attribute matrix. We develop the first non-trivial lower bound on the worst-case error of the best map from attributes to classes for this setting, even with perfect attribute detectors. The lower bound characterizes the theoretical intrinsic difficulty of the zero-shot problem based on the available information---the class-attribute matrix---and the bound is practically computable from it. Our lower bound is tight, as we show that we can always find a randomized map from attributes to classes whose expected error is upper bounded by the value of the lower bound. We show that our analysis can be predictive of how standard zero-shot methods behave in practice, including which classes will likely be confused with others.

## [Universal Rates for Interactive Learning](#)

- Steve Hanneke · Amin Karbasi · Shay Moran · Grigoris Velezgas
- abstract@[open-review](#): Consider the task of learning an unknown concept from a given concept class; to what extent does interacting with a domain expert accelerate the learning process? It is common to measure the effectiveness of learning algorithms by plotting the "learning curve", that is, the decay of the error rate as a function of the algorithm's resources (examples, queries, etc). Thus, the overarching question in this work is whether (and which kind of) interaction accelerates the learning curve. Previous work in interactive learning focused on uniform bounds on the learning rates which only capture the upper envelope of the learning curves over families of data distributions. We thus formalize our overarching question within the distribution dependent framework of universal learning, which aims to understand the performance of learning algorithms on every data distribution, but without requiring a single upper bound which applies uniformly to all distributions. Our main result reveals a fundamental trichotomy of interactive learning rates, thus providing a complete characterization of universal interactive learning. As a corollary we deduce a strong affirmative answer to our overarching question, showing that interaction is beneficial. Remarkably, we show that in important cases such benefits are realized with label queries, that is, by active learning algorithms. On the other hand, our lower bounds apply to arbitrary binary queries and, hence, they hold in any interactive learning setting.

## [How and Why to Manipulate Your Own Agent: Modeling Games between Users of Learning Agents](#)

- Yoav Kolumbus · Noam Nisan
- abstract@[open-review](#): The usage of automated learning agents is becoming increasingly prevalent in many online economic applications such as online auctions and automated trading. Motivated by such applications, this paper is dedicated to fundamental modeling and analysis of the strategic situations that the users of automated learning agents are facing. We consider strategic settings where several users engage in a repeated online interaction, assisted by regret-minimizing learning agents that repeatedly play a "game" on their behalf. We propose to view the outcomes of the agents' dynamics as inducing a "meta-game" between the users. Our main focus is on whether users can benefit in this meta-game from "manipulating" their own agents by misreporting their parameters to them. We define a general framework to model and analyze these strategic interactions between users of learning agents for general games and analyze the equilibria induced on the users in three classes of games.

## [Chain of Thought Prompting Elicits Reasoning in Large Language Models](#)

- Jason Wei · Xuezhi Wang · Dale Schuurmans · Maarten Bosma · Brian Ichter · Fei Xia · Ed Chi · Quoc V Le · Denny Zhou
- abstract@[open-review](#): We explore how generating a chain of thought---a series of intermediate reasoning steps---significantly improves the ability of large language models to perform complex reasoning. In particular, we show how such reasoning abilities emerge naturally in sufficiently large language models via a simple method called chain of thought prompting, where a few chain of thought demonstrations are provided as exemplars in prompting. Experiments on three large language models show that chain of thought prompting improves performance on a range of arithmetic, commonsense, and symbolic reasoning tasks. The empirical gains can be striking. For instance, prompting a 540B-parameter language model with just eight chain of thought exemplars achieves state of the art accuracy on the GSM8K benchmark of math word problems, surpassing even finetuned GPT-3 with a verifier.

## [Label-Aware Global Consistency for Multi-Label Learning with Single Positive Labels](#)

- Ming-Kun Xie · Jiahao Xiao · Sheng-Jun Huang
- abstract@[open-review](#): In single positive multi-label learning (SPML), only one of multiple positive labels is observed for each instance. The previous work trains the model by simply treating unobserved labels as negative ones, and designs the regularization to constrain the number of expected positive labels. However, in many real-world scenarios, the true number of positive labels is unavailable, making such methods less applicable. In this paper, we propose to solve SPML problems by designing a Label-Aware global Consistency (LAC) regularization, which leverages the manifold structure information to enhance the recovery of potential positive labels. On one hand, we first perform pseudo-labeling for each unobserved label based on its prediction probability. The consistency regularization is then imposed on model outputs to balance the fitting of identified labels and exploring of potential

positive labels. On the other hand, by enforcing label-wise embedding to maintain global consistency, LAC loss encourages the model to learn a more distinctive representation, which benefits for recovering the information of potential positive labels. Experiments on multiple benchmark datasets validate that the proposed method can achieve state-of-the-art performance for solving SPML tasks.

## [Surprise Minimizing Multi-Agent Learning with Energy-based Models](#)

- Karush Suri
- abstract@[open-review](#): Multi-Agent Reinforcement Learning (MARL) has demonstrated significant success by virtue of collaboration across agents. Recent work, on the other hand, introduces surprise which quantifies the degree of change in an agent's environment. Surprise-based learning has received significant attention in the case of single-agent entropic settings but remains an open problem for fast-paced dynamics in multi-agent scenarios. A potential alternative to address surprise may be realized through the lens of free-energy minimization. We explore surprise minimization in multi-agent learning by utilizing the free energy across all agents in a multi-agent system. A temporal Energy-Based Model (EBM) represents an estimate of surprise which is minimized over the joint agent distribution. Our formulation of the EBM is theoretically akin to the minimum conjugate entropy objective and highlights suitable convergence towards minimum surprising states. We further validate our theoretical claims in an empirical study of multi-agent tasks demanding collaboration in the presence of fast-paced dynamics. Our implementation and agent videos are available at the anonymous Project Webpage.

## [Learning Predictions for Algorithms with Predictions](#)

- Misha Khodak · Maria-Florina Balcan · Ameet Talwalkar · Sergei Vassilvitskii
- abstract@[open-review](#): A burgeoning paradigm in algorithm design is the field of algorithms with predictions, in which algorithms are designed to take advantage of a possibly-imperfect prediction of some aspect of the problem. While much work has focused on using predictions to improve competitive ratios, running times, or other performance measures, less effort has been devoted to the question of how to obtain the predictions themselves, especially in the critical online setting. We introduce a general design approach for algorithms that learn predictors: (1) identify a functional dependence of the performance measure on the prediction quality, and (2) apply techniques from online learning to learn predictors against adversarial instances, tune robustness-consistency trade-offs, and obtain new statistical guarantees. We demonstrate the effectiveness of our approach at deriving learning-algorithms by analyzing methods for bipartite matching, ski-rental, page migration, and job scheduling. In the first two settings we improve upon existing learning-theoretic results by deriving online results, obtaining better or more general statistical guarantees, and utilizing a much simpler analysis, while in the last two we provide the first learning-theoretic guarantees.

## [Fair Ranking with Noisy Protected Attributes](#)

- Anay Mehrotra · Nisheeth Vishnoi
- abstract@[open-review](#): The fair-ranking problem, which asks to rank a given set of items to maximize utility subject to group fairness constraints, has received attention in the fairness, {information retrieval}, and machine learning literature. Recent works, however, observe that errors in socially-salient (including protected) attributes of items can significantly undermine fairness guarantees of existing fair-ranking algorithms and raise the problem of mitigating the effect of such errors. We study the fair-ranking problem under a model where socially-salient attributes of items are randomly and independently perturbed. We present a fair-ranking framework that incorporates group fairness requirements along with probabilistic information about perturbations in socially-salient attributes. We provide provable guarantees on the fairness and utility attainable by our framework and show that it is information-theoretically impossible to significantly beat these guarantees. Our framework works for multiple non-disjoint attributes and a general class of fairness constraints that includes proportional and equal representation. Empirically, we observe that, compared to baselines, our algorithm outputs rankings with higher fairness, and has a similar or better fairness-utility trade-off compared to baselines.

## [Semi-Supervised Video Salient Object Detection Based on Uncertainty-Guided Pseudo Labels](#)

- chenyang lu · Yongri Piao · Miao Zhang · Huchuan Lu
- abstract@[open-review](#): Semi-Supervised Video Salient Object Detection (SS-VSOD) is challenging because of the lack of temporal information in video sequences caused by sparse annotations. Most works address this problem by generating pseudo labels for unlabeled data. However, error-prone pseudo labels negatively affect the VOSD model. Therefore, a deeper insight into pseudo labels should be developed. In this work, we aim to explore 1) how to utilize the incorrect predictions in pseudo labels to guide the network to generate more robust pseudo labels and 2) how to further screen out the noise that still exists in the improved pseudo labels. To this end, we propose an Uncertainty-Guided Pseudo Label Generator (UGPLG), which makes full use of inter-frame information to ensure the temporal consistency of the pseudo labels and improves the robustness of the pseudo labels by strengthening the learning of difficult scenarios. Furthermore, we also introduce the adversarial learning to address the noise problems in pseudo labels, guaranteeing the positive guidance of pseudo labels during model training. Experimental results demonstrate that our methods outperform existing semi-supervised method and partial fully-supervised methods across five public benchmarks of DAVIS, FBMS, MCL, ViSal and SegTrack-V2.

## [Learning Expressive Meta-Representations with Mixture of Expert Neural Processes](#)

- Qi Wang · Herke van Hoof
- abstract@[open-review](#): Neural processes (NPs) formulate exchangeable stochastic processes and are promising models for meta learning that do not require gradient updates during the testing phase. However, most NP variants place a strong emphasis on a global latent variable. This weakens the approximation power and restricts the scope of applications using NP variants, especially when data generative processes are complicated. To resolve these issues, we propose to combine the Mixture of Expert models with Neural Processes to develop more expressive exchangeable stochastic processes, referred to as Mixture of Expert Neural Processes (MoE-NPs). Then we apply MoE-NPs to both few-shot supervised learning and meta reinforcement learning tasks. Empirical results demonstrate MoE-NPs' strong generalization capability to unseen tasks in these benchmarks.

## [Improving GANs with A Dynamic Discriminator](#)

- Ceyuan Yang · Yujun Shen · Yinghao Xu · Deli Zhao · Bo Dai · Bolei Zhou
- abstract@[open-review](#): Discriminator plays a vital role in training generative adversarial networks (GANs) via distinguishing real and synthesized samples. While the real data distribution remains the same, the synthesis distribution keeps varying because of the evolving generator, and thus effects a corresponding change of the bi-classification task assigned to the discriminator. We argue that a discriminator with an on-the-fly adjustment on its capacity can better accommodate such a time-varying task. A comprehensive empirical study confirms that the proposed training strategy, termed as DynamicD, improves the synthesis performance without incurring any additional computation cost or training objectives. Two capacity adjusting schemes are developed for training GANs under different data regimes: i) given a sufficient amount of training data, the discriminator benefits from a progressively increased learning capacity, and ii) when the training data is limited, gradually decreasing the layer width mitigates the over-fitting issue of the discriminator. Experiments on both 2D and 3D-aware image synthesis tasks conducted on a range of datasets substantiate the generalizability of our DynamicD as well as its substantial improvement over the baselines. Furthermore, DynamicD is synergistic to other discriminator-improving approaches (including data augmentation, regularizers, and pre-training), and brings continuous performance gain when combined with them for learning GANs. Code will be made publicly available.

## QC-StyleGAN - Quality Controllable Image Generation and Manipulation

- Dat Viet Thanh Nguyen Â· Phong Tran The Â· Tan M. Dinh Â· Cuong Pham Â· Anh Tran
- abstract@[open-review](#): The introduction of high-quality image generation models, particularly the StyleGAN family, provides a powerful tool to synthesize and manipulate images. However, existing models are built upon high-quality (HQ) data as desired outputs, making them unfit for in-the-wild low-quality (LQ) images, which are common inputs for manipulation. In this work, we bridge this gap by proposing a novel GAN structure that allows for generating images with controllable quality. The network can synthesize various image degradation and restore the sharp image via a quality control code. Our proposed QC-StyleGAN can directly edit LQ images without altering their quality by applying GAN inversion and manipulation techniques. It also provides for free an image restoration solution that can handle various degradations, including noise, blur, compression artifacts, and their mixtures. Finally, we demonstrate numerous other applications such as image degradation synthesis, transfer, and interpolation.

## Pseudo-Riemannian Graph Convolutional Networks

- Bo Xiong Â· Shichao Zhu Â· Nico Potyka Â· Shirui Pan Â· Chuan Zhou Â· Steffen Staab
- abstract@[open-review](#): Graph Convolutional Networks (GCNs) are powerful frameworks for learning embeddings of graph-structured data. GCNs are traditionally studied through the lens of Euclidean geometry. Recent works find that non-Euclidean Riemannian manifolds provide specific inductive biases for embedding hierarchical or spherical data. However, they cannot align well with data of mixed graph topologies. We consider a larger class of pseudo-Riemannian manifolds that generalize hyperboloid and sphere. We develop new geodesic tools that allow for extending neural network operations into geodesically disconnected pseudo-Riemannian manifolds. As a consequence, we derive a pseudo-Riemannian GCN that models data in pseudo-Riemannian manifolds of constant nonzero curvature in the context of graph neural networks. Our method provides a geometric inductive bias that is sufficiently flexible to model mixed heterogeneous topologies like hierarchical graphs with cycles. We demonstrate the representational capabilities of this method by applying it to the tasks of graph reconstruction, node classification, and link prediction on a series of standard graphs with mixed topologies. Empirical results demonstrate that our method outperforms Riemannian counterparts when embedding graphs of complex topologies.

## NeMF: Neural Motion Fields for Kinematic Animation

- Chengan He Â· Jun Saito Â· James Zachary Â· Holly Rushmeier Â· Yi Zhou
- abstract@[open-review](#): We present an implicit neural representation to learn the spatio-temporal space of kinematic motions. Unlike previous work that represents motion as discrete sequential samples, we propose to express the vast motion space as a continuous function over time, hence the name Neural Motion Fields (NeMF). Specifically, we use a neural network to learn this function for miscellaneous sets of motions, which is designed to be a generative model conditioned on a temporal coordinate  $t$  and a random vector  $z$  for controlling the style. The model is then trained as a Variational Autoencoder (VAE) with motion encoders to sample the latent space. We train our model with diverse human motion dataset and quadruped dataset to prove its versatility, and finally deploy it as a generic motion prior to solve task-agnostic problems and show its superiority in different motion generation and editing applications, such as motion interpolation, in-betweening, and re-navigating.

## Rethinking and Improving Robustness of Convolutional Neural Networks: a Shapley Value-based Approach in Frequency Domain

- Yiting Chen Â· Qibing Ren Â· Junchi Yan
- abstract@[open-review](#): The existence of adversarial examples poses concerns for the robustness of convolutional neural networks (CNN), for which a popular hypothesis is about the frequency bias phenomenon: CNNs rely more on high-frequency components (HFC) for classification than humans, which causes the brittleness of CNNs. However, most previous works manually select and roughly divide the image frequency spectrum and conduct qualitative analysis. In this work, we introduce Shapley value, a metric of cooperative game theory, into the frequency domain and propose to quantify the positive (negative) impact of every frequency component of data on CNNs. Based on the Shapley value, we quantify the impact in a fine-grained way and show intriguing instance disparity. Statistically, we investigate adversarial training(AT) and the adversarial attack in the frequency domain. The observations motivate us to perform an in-depth analysis and lead to multiple novel hypotheses about i) the cause of adversarial robustness of the AT model; ii) the fairness problem of AT between different classes in the same dataset; iii) the attack bias on different frequency components. Finally, we propose a Shapley-value guided data augmentation technique for improving the robustness. Experimental results on image classification benchmarks show its effectiveness.

## On the Convergence of Stochastic Multi-Objective Gradient Alteration and Beyond

- Shiji Zhou Â· Wenpeng Zhang Â· Jiyan Jiang Â· Wenliang Zhong Â· Jinjie GU Â· Wenwu Zhu
- abstract@[open-review](#): The conflicting gradients problem is the major bottleneck of the effective training of models that deal with multiple objectives. To resolve this problem, several gradient alteration techniques, such as PCGrad, MGDA, and CAGrad, are developed by directly altering the conflicting gradients to refined ones with fewer or even no conflicts. However, the design and analysis of these techniques are mainly conducted under the exact full-batch gradient setting, ignoring that they are primarily applied with stochastic mini-batch gradients. In this paper, we point out that the stochastic gradient alteration algorithms may fail to converge to Pareto optimal solutions. We summarize these seemingly different algorithms into a unified algorithmic framework, where the descent direction is given by the composition of the gradients w.r.t. the multiple objectives. Then we provide a simple two-objective convex optimization instance to illustrate the non-convergence issue in detail, which shows that the non-convergence results from the determination of the composite weights solely by the stochastic gradients. To fix this issue, we propose a novel composite weights determination scheme that exponentially averages the past calculated weights. Finally, we show that the resulting new variants of stochastic gradient alteration converge to Pareto optimality under the unified framework and also give rise to improved empirical performance for PCGrad, MGDA, and CAGrad.

## Causality-driven Hierarchical Structure Discovery for Reinforcement Learning

- shaohui peng Â· Xing Hu Â· Rui Zhang Â· Ke Tang Â· Jiaming Guo Â· Qi Yi Â· Ruizhi Chen Â· xishan zhang Â· Zidong Du Â· Ling Li Â· Qi Guo Â· Yunji Chen
- abstract@[open-review](#): Hierarchical reinforcement learning (HRL) has been proven to be effective for tasks with sparse rewards, for it can improve the agent's exploration efficiency by discovering high-quality hierarchical structures (e.g., subgoals or options). However, automatically discovering high-quality hierarchical structures is still a great challenge. Previous HRL methods can only find the hierarchical structures in simple environments, as they are mainly achieved through the randomness of agent's policies during exploration. In complicated environments, such a randomness-driven exploration paradigm can hardly discover high-quality hierarchical structures because of the low exploration efficiency. In this paper, we propose CDHRL, a causality-driven hierarchical reinforcement learning framework, to build high-quality hierarchical structures efficiently in complicated environments. The key insight is that the causalities among environment variables are naturally fit for modeling reachable subgoals and their dependencies; thus, the causality is suitable to be the guidance in building high-quality hierarchical structures. Roughly, we build the hierarchy of subgoals based on causality autonomously, and utilize the subgoal-based policies to unfold further causality efficiently. Therefore, CDHRL leverages a causality-driven discovery instead of a randomness-driven exploration for high-quality hierarchical structure construction. The results in two complex environments, 2D-Minecraft and Eden, show that CDHRL can discover high-quality hierarchical structures and significantly enhance exploration efficiency.

## A Closer Look at Offline RL Agents

- Yuwei Fu · Di Wu · Benoit Boulet
- abstract@[open-review](#): Despite recent advances in the field of Offline Reinforcement Learning (RL), less attention has been paid to understanding the behaviors of learned RL agents. As a result, there remain some gaps in our understandings, i.e., why is one offline RL agent more performant than another? In this work, we first introduce a set of experiments to evaluate offline RL agents, focusing on three fundamental aspects: representations, value functions and policies. Counterintuitively, we show that a more performant offline RL agent can learn relatively low-quality representations and inaccurate value functions. Furthermore, we showcase that the proposed experiment setups can be effectively used to diagnose the bottleneck of offline RL agents. Inspired by the evaluation results, a novel offline RL algorithm is proposed by a simple modification of IQL and achieves SOTA performance. Finally, we investigate when a learned dynamics model is helpful to model-free offline RL agents, and introduce an uncertainty-based sample selection method to mitigate the problem of model noises. Code is available at: <https://anonymous.4open.science/r/RIQL-BE73>.

## [Picking on the Same Person: Does Algorithmic Monoculture lead to Outcome Homogenization?](#)

- Rishi Bommasani · Kathleen A. Creel · Ananya Kumar · Dan Jurafsky · Percy Liang
- abstract@[open-review](#): As the scope of machine learning broadens, we observe a recurring theme of  $\text{algorithmic monoculture}$ : the same systems, or systems that share components (e.g., datasets, models), are deployed by multiple decision-makers. While sharing offers advantages like amortizing effort, it also has risks. We introduce and formalize one such risk,  $\text{outcome homogenization}$ , defined here as the extent to which particular individuals or groups experience the same outcomes across different deployments. If the same individuals or groups exclusively experience undesirable outcomes, this may institutionalize systemic exclusion and reinscribe social hierarchy. We relate algorithmic monoculture and outcome homogenization by proposing the  $\text{component sharing hypothesis}$ : if algorithmic systems are increasingly built on the same data or models, then they will increasingly homogenize outcomes. We test this hypothesis on algorithmic fairness benchmarks based on the US Census, demonstrating that increased data-sharing exacerbates homogenization, especially for small datasets. Further, given the current regime in AI of foundation models, i.e., pretrained models that can be adapted to myriad downstream tasks, we test whether model-sharing homogenizes outcomes across tasks. We observe mixed results: we find that for both vision and language settings, the specific methods for adapting a foundation model significantly influence the degree of outcome homogenization. We also identify societal challenges that inhibit the measurement, diagnosis, and rectification of outcome homogenization in deployed machine learning systems.

## [Robust Imitation via Mirror Descent Inverse Reinforcement Learning](#)

- Dong-Sig Han · Hyunseo Kim · Hyundo Lee · JeHwan Ryu · Byoung-Tak Zhang
- abstract@[open-review](#): Recently, adversarial imitation learning has shown a scalable reward acquisition method for inverse reinforcement learning (IRL) problems. However, estimated reward signals often become uncertain and fail to train a reliable statistical model since the existing methods tend to solve hard optimization problems directly. Inspired by a first-order optimization method called mirror descent, this paper proposes to predict a sequence of reward functions, which are iterative solutions for a constrained convex problem. IRL solutions derived by mirror descent are tolerant to the uncertainty incurred by target density estimation since the amount of reward learning is regulated with respect to local geometric constraints. We prove that the proposed mirror descent update rule ensures robust minimization of a Bregman divergence in terms of a rigorous regret bound of  $\mathcal{O}(1/T)$  for step sizes  $\{\eta_t\}_{t=1}^T$ . Our IRL method was applied on top of an adversarial framework, and it outperformed existing adversarial methods in an extensive suite of benchmarks.

## [A Probabilistic Graph Coupling View of Dimension Reduction](#)

- Hugues Van Assel · Thibault Espinasse · Julien Chiquet · Franck Picard
- abstract@[open-review](#): Most popular dimension reduction (DR) methods like t-SNE and UMAP are based on minimizing a cost between input and latent pairwise similarities. Though widely used, these approaches lack clear probabilistic foundations to enable a full understanding of their properties and limitations. To that extent, we introduce a unifying statistical framework based on the coupling of hidden graphs using cross entropy. These graphs induce a Markov random field dependency structure among the observations in both input and latent spaces. We show that existing pairwise similarity DR methods can be retrieved from our framework with particular choices of priors for the graphs. Moreover this reveals that these methods relying on shift-invariant kernels suffer from a statistical degeneracy that explains poor performances in conserving coarse-grain dependencies. New links are drawn with PCA which appears as a non-degenerate graph coupling model.

## [DTMD: Learning Improvement of Spiking Neural Networks with Dynamic Thresholding Neurons and Moderate Dropout](#)

- SIQI WANG · Tee Hiang Cheng · Meng-Hiot Lim
- abstract@[open-review](#): Spiking Neural Networks (SNNs) have shown great promise in processing spatio-temporal data, mimicking biological neuronal mechanisms, and saving computational power. However, most SNNs use fixed model regardless of their locations in the network. This limits SNNs' capability of transmitting precise information in the network, which becomes worse for deeper SNNs. Some researchers try to use specified parametric models in different network layers or regions, but most still use preset or suboptimal parameters. Inspired by the neuroscience observation that different neuronal mechanisms exist in disparate brain regions, we propose a new spiking neuronal mechanism, named dynamic thresholding, to address this issue. Utilizing learnable threshold values, dynamic thresholding enables flexible neuronal mechanisms across layers, proper information flow within the network, and fast network convergence. In addition, we propose a moderate dropout method to serve as an enhancement technique to minimize inconsistencies between independent dropout runs. Finally, we evaluate the robustness of the proposed dynamic thresholding and moderate dropout for image classification with different initial thresholds for various types of datasets. Our proposed methods produce superior results compared to other approaches for almost all datasets with fewer timesteps.

## [Communication-Efficient Topologies for Decentralized Learning with \$\mathcal{O}\(1\)\$ Consensus Rate](#)

- Zuoqing Song · Weijian Li · Kexin Jin · Lei Shi · Ming Yan · Wotao Yin · Kun Yuan
- abstract@[open-review](#): Decentralized optimization is an emerging paradigm in distributed learning in which agents achieve network-wide solutions by peer-to-peer communication without the central server. Since communication tends to be slower than computation, when each agent communicates with only a few neighboring agents per iteration, they can complete iterations faster than with more agents or a central server. However, the total number of iterations to reach a network-wide solution is affected by the speed at which the information of the agents is ``mixed'' by communication. We found that popular communication topologies either have large degrees (such as stars and complete graphs) or are ineffective at mixing information (such as rings and grids). To address this problem, we propose a new family of topologies, EquiTopo, which has an (almost) constant degree and network-size-independent consensus rate which is used to measure the mixing efficiency. In the proposed family, EquiStatic has a degree of  $\Theta(\ln(n))$ , where  $n$  is the network size, and a series of time-varying one-peer topologies, EquiDyn, has a constant degree of 1. We generate EquiDyn through a certain random sampling procedure. Both of them achieve  $n$ -independent consensus rate. We apply them to decentralized SGD and decentralized gradient tracking and obtain faster communication and better convergence, both theoretically and empirically.

## [Semantic Field of Words Represented as Non-Linear Potential Functions](#)

- Xin Du · Kumiko Tanaka-Ishii
- abstract@[open-review](#): State-of-the-art word embeddings presume a linear vector space, but this approach does not easily incorporate the nonlinearity that is necessary to represent polysemy. We thus propose a novel semantic Field REpresentation, called FIRE, which is a  $D$ -dimensional field in which

every word is represented as a set of its locations and a nonlinear function covering the field. The strength of a word's relation to another word at a certain location is measured as the function value at that location. With FIRE, compositionality is represented via functional additivity, whereas polysemy is represented via the set of points and the function's multimodality. By implementing FIRE for English and comparing it with previous representation methods via word and sentence similarity tasks, we show that FIRE produces comparable or even better results. In an evaluation of polysemy to predict the number of word senses, FIRE greatly outperformed BERT and Word2vec, providing evidence of how FIRE represents polysemy.

## [Domain Generalization by Learning and Removing Domain-specific Features](#)

- Yu Ding Â· Lei Wang Â· Bin Liang Â· Shuming Liang Â· Yang Wang Â· Fang Chen
- abstract@[open-review](#): Deep Neural Networks (DNNs) suffer from domain shift when the test dataset follows a distribution different from the training dataset. Domain generalization aims to tackle this issue through learning a model that can generalize to unseen domains. In this paper, we propose a new approach that aims to explicitly remove domain-specific features for domain generalization. Following this approach, we propose a novel framework called Learning and Removing Domain-specific features for Generalization (LRDG) that learns a domain-invariant model by tactically removing domain-specific features from the input images. Specifically, we design a classifier to effectively learn domain-specific features for each source domain, respectively. We then develop an encoder-decoder network to map each input image into a new image space where the learned domain-specific features are removed. With the images output by the encoder-decoder network, another classifier is designed to learn the domain-invariant features to conduct image classification. Extensive experiments demonstrate that our framework achieves superior performance compared with the state-of-the-art methods.

## [Functional Ensemble Distillation](#)

- Coby Penso Â· Idan Achituvé Â· Ethan Fetaya
- abstract@[open-review](#): Bayesian models have many desirable properties, most notable is their ability to generalize from limited data and to properly estimate the uncertainty in their predictions. However, these benefits come at a steep computational cost as Bayesian inference, in most cases, is computationally intractable. One popular approach to alleviate this problem is using a Monte-Carlo estimation with an ensemble of models sampled from the posterior. However, this approach still comes at a significant computational cost, as one needs to store and run multiple models at test time. In this work, we investigate how to best distill an ensemble's predictions using an efficient model. First, we argue that current approaches are limited as they are constrained to classification and the Dirichlet distribution. Second, in many limited data settings, all ensemble members achieve nearly zero training loss, namely, they produce near-identical predictions on the training set which results in sub-optimal distilled models. To address both problems, we propose a novel and general distillation approach, named Functional Ensemble Distillation (FED), and we investigate how to best distill an ensemble in this setting. We find that learning the distilled model via a simple augmentation scheme in the form of mixup augmentation significantly boosts the performance. We evaluated our method on several tasks and showed that it achieves superior results in both accuracy and uncertainty estimation compared to current approaches.

## [PALBERT: Teaching ALBERT to Ponder](#)

- Nikita Balagansky Â· Daniil Gavrilov
- abstract@[open-review](#): Currently, pre-trained models can be considered the default choice for a wide range of NLP tasks. Despite their SoTA results, there is practical evidence that these models may require a different number of computing layers for different input sequences, since evaluating all layers leads to overconfidence in wrong predictions (namely overthinking). This problem can potentially be solved by implementing adaptive computation time approaches, which were first designed to improve inference speed. Recently proposed PonderNet may be a promising solution for performing an early exit by treating the exit layer's index as a latent variable. However, the originally proposed exit criterion, relying on sampling from trained posterior distribution on the probability of exiting from the  $i$ -th layer, introduces major variance in exit layer indices, significantly reducing the resulting model's performance. In this paper, we propose improving PonderNet with a novel deterministic Q-exit criterion and a revisited model architecture. We adapted the proposed mechanism to ALBERT and RoBERTa and compared it with recent methods for performing an early exit. We observed that the proposed changes can be considered significant improvements on the original PonderNet architecture and outperform PABEE on a wide range of GLUE tasks. In addition, we also performed an in-depth ablation study of the proposed architecture to further understand Lambda layers and their performance.

## [Provable Generalization of Overparameterized Meta-learning Trained with SGD](#)

- Yu Huang Â· Yingbin Liang Â· Longbo Huang
- abstract@[open-review](#): Despite the empirical success of deep meta-learning, theoretical understanding of overparameterized meta-learning is still limited. This paper studies the generalization of a widely used meta-learning approach, Model-Agnostic Meta-Learning (MAML), which aims to find a good initialization for fast adaptation to new tasks. Under a mixed linear regression model, we analyze the generalization properties of MAML trained with SGD in the overparameterized regime. We provide both upper and lower bounds for the excess risk of MAML, which captures how SGD dynamics affect these generalization bounds. With such sharp characterizations, we further explore how various learning parameters impact the generalization capability of overparameterized MAML, including explicitly identifying typical data and task distributions that can achieve diminishing generalization error with overparameterization, and characterizing the impact of adaptation learning rate on both excess risk and the early stopping time. Our theoretical findings are further validated by experiments.

## [Monte Carlo Tree Descent for Black-Box Optimization](#)

- Yaoguang Zhai Â· Sicun Gao
- abstract@[open-review](#): The key to black-box optimization is the efficient search among regions with widely-varying numerical properties to achieve low-regret descent. Monte Carlo Tree Search (MCTS) methods have recently been introduced to improve Bayesian optimization by computing partitioning of the search space and balancing exploration and exploitation. Extending this promising framework, we study how to better balance sample-driven descent and Bayesian optimization for faster descent with fewer samples. At the vertices of the search trees, we first introduce new descent methods that incorporate stochastic and direct search. We then design new ways of balancing progress and uncertainty, and propose new branch selection, tree expansion, and backpropagation policies. Overall, the proposed MCTS puts more emphasis on sampling for faster descent, and uses localized Gaussian Processes as auxiliary metrics in both exploitation and exploration. We show experimentally that the new designs can achieve good optimization results compared to state-of-the-art methods on challenging benchmark problems.

## [Pyramid Attention For Source Code Summarization](#)

- Lei Chai Â· Ming LI
- abstract@[open-review](#): In this paper, we present a multi-granularity method for the task of source code summarization, which generates a concise functional description for the given code snippet. We notice that skilled programmers write and read source codes in a hierarchical way and pay close attention to the conceptual entities like statements, tokens, sub-tokens, and the mapping relations between them. The entities have specific emphasis according to their granularities, e.g., statements in coarse-granularity reveal the global logical semantics of code, and the sub-tokens in fine-granularity are more related to the textual semantics. Driven by this observation, we argue that a multi-granularities formulation incorporating the entities in different granularities may benefit the code summarization task. Given a code snippet, we first construct a pyramid-shaped input representation, and a pyramid attention mechanism is proposed for efficient feature aggregation and distribution across different hierarchies. We instantiate our multi-granularity method

using the proposed pyramid attention and name it PA-former (Pyramid Attention Transformer), which is evaluated on two source code summarization benchmarks where it surpasses the prior works and achieves new state-of-the-art results.

## [Reconstruction on Trees and Low-Degree Polynomials](#)

- Frederic Koehler · Elchanan Mossel
- abstract@[open-review](#): The study of Markov processes and broadcasting on trees has deep connections to a variety of areas including statistical physics, graphical models, phylogenetic reconstruction, MCMC algorithms, and community detection in random graphs. Notably, the celebrated Belief Propagation (BP) algorithm achieves Bayes-optimal performance for the reconstruction problem of predicting the value of the Markov process at the root of the tree from its values at the leaves. Recently, the analysis of low-degree polynomials has emerged as a valuable tool for predicting computational-to-statistical gaps. In this work, we investigate the performance of low-degree polynomials for the reconstruction problem on trees. Perhaps surprisingly, we show that there are simple tree models with  $N$  leaves and bounded arity where (1) nontrivial reconstruction of the root value is possible with a simple polynomial time algorithm and with robustness to noise, but not with any polynomial of degree  $N^{c}$  for  $c > 0$  a constant depending only on the arity, and (2) when the tree is unknown and given multiple samples with correlated root assignments, nontrivial reconstruction of the root value is possible with a simple Statistical Query algorithm but not with any polynomial of degree  $N^c$ . These results clarify some of the limitations of low-degree polynomials vs. polynomial time algorithms for Bayesian estimation problems. They also complement recent work of Moitra, Mossel, and Sandon who studied the circuit complexity of Belief Propagation. As a consequence of our main result, we are able to prove a result of independent interest regarding the performance of RBF kernel ridge regression for learning to predict the root coloration: for some  $c' > 0$  depending only on the arity,  $\exp(N^{c'})$  many samples are needed for the kernel regression to obtain nontrivial correlation with the true regression function (BP). We pose related open questions about low-degree polynomials and the Kesten-Stigum threshold.

## [Neural Collapse with Normalized Features: A Geometric Analysis over the Riemannian Manifold](#)

- Can Yaras · Peng Wang · Zhihui Zhu · Laura Balzano · Qing Qu
- abstract@[open-review](#): When training overparameterized deep networks for classification tasks, it has been widely observed that the learned features exhibit a so-called "neural collapse" phenomenon. More specifically, for the output features of the penultimate layer, for each class the within-class features converge to their means, and the means of different classes exhibit a certain tight frame structure, which is also aligned with the last layer's classifier. As feature normalization in the last layer becomes a common practice in modern representation learning, in this work we theoretically justify the neural collapse phenomenon for normalized features. Based on an unconstrained feature model, we simplify the empirical loss function in a multi-class classification task and obtain a nonconvex optimization problem over the Riemannian manifold by constraining all features and classifiers over the sphere. In this context, we analyze the nonconvex landscape of the Riemannian optimization problem over the product of spheres, showing a benign global landscape in the sense that the only global minimizers are the neural collapse solutions while all other critical points are strict saddles with negative curvature. Experimental results on practical deep networks corroborate our theory and demonstrate that better representations can be learned faster via feature normalization.

## [Multiclass Learnability Beyond the PAC Framework: Universal Rates and Partial Concept Classes](#)

- Alkis Kalavasis · Grigorios Velezgas · Amin Karbasi
- abstract@[open-review](#): In this paper we study the problem of multiclass classification with a bounded number of different labels  $k$ , in the realizable setting. We extend the traditional PAC model to a) distribution-dependent learning rates, and b) learning rates under data-dependent assumptions. First, we consider the universal learning setting (Bousquet, Hanneke, Moran, van Handel and Yehudayoff, STOC'21), for which we provide a complete characterization of the achievable learning rates that holds for every fixed distribution. In particular, we show the following trichotomy: for any concept class, the optimal learning rate is either exponential, linear or arbitrarily slow. Additionally, we provide complexity measures of the underlying hypothesis class that characterize when these rates occur. Second, we consider the problem of multiclass classification with structured data (such as data lying on a low dimensional manifold or satisfying margin conditions), a setting which is captured by partial concept classes (Alon, Hanneke, Holzman and Moran, FOCS'21). Partial concepts are functions that can be undefined in certain parts of the input space. We extend the traditional PAC learnability of total concept classes to partial concept classes in the multiclass setting and investigate differences between partial and total concepts.

## [Learning from Stochastically Revealed Preference](#)

- John Birge · Xiaocheng Li · Chunlin Sun
- abstract@[open-review](#): We study the learning problem of revealed preference in a stochastic setting: a learner observes the utility-maximizing actions of a set of agents whose utility follows some unknown distribution, and the learner aims to infer the distribution through the observations of actions. The problem can be viewed as a single-constraint special case of the inverse linear optimization problem. Existing works all assume that all the agents share one common utility which can easily be violated under practical contexts. In this paper, we consider two settings for the underlying utility distribution: a Gaussian setting where the customer utility follows the von Mises-Fisher distribution, and a  $\delta$ -corruption setting where the customer utility distribution concentrates on one fixed vector with high probability and is arbitrarily corrupted otherwise. We devise Bayesian approaches for parameter estimation and develop theoretical guarantees for the recovery of the true parameter. We illustrate the algorithm performance through numerical experiments.

## [A Single-timescale Analysis for Stochastic Approximation with Multiple Coupled Sequences](#)

- Han Shen · Tianyi Chen
- abstract@[open-review](#): Stochastic approximation (SA) with multiple coupled sequences has found broad applications in machine learning such as bilevel learning and reinforcement learning (RL). In this paper, we study the finite-time convergence of nonlinear SA with multiple coupled sequences. Different from existing multi-timescale analysis, we seek scenarios where a fine-grained analysis can provide a tight performance guarantee for single-timescale multi-sequence SA (STSA). At the heart of our analysis is the smoothness property of the fixed points in multi-sequence SA that holds in many applications. When all sequences have strongly monotone increments, we establish the iteration complexity of  $O(\epsilon^{-1})$  to achieve  $\epsilon$ -accuracy, which improves the existing  $O(\epsilon^{-1.5})$  complexity for two coupled sequences. When the main sequence does not have a strongly monotone increment, we establish the iteration complexity of  $O(\epsilon^{-2})$ . We showcase the power of our result by applying it to stochastic bilevel and compositional optimization problems, as well as RL problems, all of which recover the best known or lead to improvements over their existing guarantees.

## [Context-Based Dynamic Pricing with Partially Linear Demand Model](#)

- Jinzhi Bu · David Simchi-Levi · Chonghuan Wang
- abstract@[open-review](#): In today's data-rich environment, context-based dynamic pricing has gained much attention. To model the demand as a function of price and context, the existing literature either adopts a parametric model or a non-parametric model. The former is easier to implement but may suffer from model mis-specification, whereas the latter is more robust but does not leverage many structural properties of the underlying problem. This paper combines these two approaches by studying the context-based dynamic pricing with online learning, where the unknown expected demand admits a semi-parametric partially linear structure. Specifically, we consider two demand models, whose expected demand at price  $p$  and context  $x$  is given by  $bp+g(x)$  and  $f(p)+a^\top x$  respectively. We assume that  $g(x)$  is  $\beta$ -Hölder continuous in the first model, and

$f(p)$  is  $k$ -th-order smooth with an additional parameter  $\delta$  in the second model. For both models, we design an efficient online learning algorithm with provable regret upper bounds, and establish matching lower bounds. This enables us to characterize the statistical complexity for the two learning models, whose optimal regret rates are  $\tilde{\Theta}(\sqrt{T} \vee T^{\frac{d}{d+2\delta}})$  and  $\tilde{\Theta}(\sqrt{T} \vee (\delta T^{k+1})^{\frac{1}{2k+1}})$  respectively. The numerical results demonstrate that our learning algorithms are more effective than benchmark algorithms, and also reveal the effects of parameters  $d$ ,  $\beta$  and  $\delta$  on the algorithm's empirical regret, which are consistent with our theoretical findings.

## [Stochastic Second-Order Methods Provably Beat SGD For Gradient-Dominated Functions](#)

- Mohammadsaeed Masiha · Saber Salehkaleybar · Niao He · Negar Kiyavash · Patrick Thiran
- abstract@[open-review](#): We study the performance of Stochastic Cubic Regularized Newton (SCRN) on a class of functions satisfying gradient dominance property which holds in a wide range of applications in machine learning and signal processing. This condition ensures that any first-order stationary point is a global optimum. We prove that SCRN improves the best-known sample complexity of stochastic gradient descent in achieving  $\epsilon$ -global optimum by a factor of  $\mathcal{O}(\epsilon^{-1/2})$ . Even under a weak version of gradient dominance property, which is applicable to policy-based reinforcement learning (RL), SCRN achieves the same improvement over stochastic policy gradient methods. Additionally, we show that the sample complexity of SCRN can be improved by a factor of  $\mathcal{O}(\epsilon^{-1/2})$  using a variance reduction method with time-varying batch sizes. Experimental results in various RL settings showcase the remarkable performance of SCRN compared to first-order methods.

## [Direct Advantage Estimation](#)

- Hsiao-Ru Pan · Nico Grärtler · Alexander Neitz · Bernhard Schölkopf
- abstract@[open-review](#): The predominant approach in reinforcement learning is to assign credit to actions based on the expected return. However, we show that the return may depend on the policy in a way which could lead to excessive variance in value estimation and slow down learning. Instead, we show that the advantage function can be interpreted as causal effects and shares similar properties with causal representations. Based on this insight, we propose Direct Advantage Estimation (DAE), a novel method that can model the advantage function and estimate it directly from on-policy data while simultaneously minimizing the variance of the return without requiring the (action-)value function. We also relate our method to Temporal Difference methods by showing how value functions can be seamlessly integrated into DAE. The proposed method is easy to implement and can be readily adapted by modern actor-critic methods. We evaluate DAE empirically on three discrete control domains and show that it can outperform generalized advantage estimation (GAE), a strong baseline for advantage estimation, on a majority of the environments when applied to policy optimization.

## [Towards Effective Multi-Modal Interchanges in Zero-Resource Sounding Object Localization](#)

- Yang Zhao · Chen Zhang · Haifeng Huang · Haoyuan Li · Zhou Zhao
- abstract@[open-review](#): Aiming to locate the object that emits a specified sound in complex scenes, the task of sounding object localization bridges two perception-oriented modalities of vision and acoustics, and brings enormous research value to the comprehensive perceptual understanding of machine intelligence. Although there are massive training data collected in this field, few of them contain accurate bounding box annotations, hindering the learning process and further application of proposed models. In order to address this problem, we try to explore an effective multi-modal knowledge transfer strategy to obtain precise knowledge from other similar tasks and transfer it through well-aligned multi-modal data to deal with this task in a zero-resource manner. Concretely, we design and propose a novel Two-stream Universal Referring localization Network (TURN), which is composed of a localization stream and an alignment stream to carry out different functions. The former is utilized to extract the knowledge related to referring object localization from the image grounding task, while the latter is devised to learn a universal semantic space shared between texts and audios. Moreover, we further develop an adaptive sampling strategy to automatically identify the overlap between different data domains, thus boosting the performance and stability of our model. The extensive experiments on various publicly-available benchmarks demonstrate that TURN can achieve competitive performance compared with the state-of-the-art approaches without using any data in this field, which verifies the feasibility of our proposed mechanisms and strategies.

## [OrdinalCLIP: Learning Rank Prompts for Language-Guided Ordinal Regression](#)

- Wanhua Li · Xiaoke Huang · Zheng Zhu · Yansong Tang · Xiu Li · Jie Zhou · Jiwen Lu
- abstract@[open-review](#): This paper presents a language-powered paradigm for ordinal regression. Existing methods usually treat each rank as a category and employ a set of weights to learn these concepts. These methods are easy to overfit and usually attain unsatisfactory performance as the learned concepts are mainly derived from the training set. Recent large pre-trained vision-language models like CLIP have shown impressive performance on various visual tasks. In this paper, we propose to learn the rank concepts from the rich semantic CLIP latent space. Specifically, we reformulate this task as an image-language matching problem with a contrastive objective, which regards labels as text and obtains a language prototype from a text encoder for each rank. While prompt engineering for CLIP is extremely time-consuming, we propose OrdinalCLIP, a differentiable prompting method for adapting CLIP for ordinal regression. OrdinalCLIP consists of learnable context tokens and learnable rank embeddings. The learnable rank embeddings are constructed by explicitly modeling numerical continuity, resulting in well-ordered, compact language prototypes in the CLIP space. Once learned, we can only save the language prototypes and discard the huge language model, resulting in zero additional computational overhead compared with the linear head counterpart. Experimental results show that our paradigm achieves competitive performance in general ordinal regression tasks, and gains improvements in few-shot and distribution shift settings for age estimation. The code is available at <https://github.com/xk-huang/OrdinalCLIP>.

## [Equivariant Representation in Recurrent Networks with a Continuous Manifold of Attractors](#)

- Wenhao Zhang · Ying Nian Wu · Si Wu
- abstract@[open-review](#): Equivariant representation is necessary for the brain and artificial perceptual systems to faithfully represent the stimulus under some (Lie) group transformations. However, it remains unknown how recurrent neural circuits in the brain represent the stimulus equivariantly, nor the neural representation of abstract group operators. The present study uses a one-dimensional (1D) translation group as an example to explore the general recurrent neural circuit mechanism of the equivariant stimulus representation. We found that a continuous attractor network (CAN), a canonical neural circuit model, self-consistently generates a continuous family of stationary population responses (attractors) that represents the stimulus equivariantly. Inspired by the Drosophila's compass circuit, we found that the 1D translation operators can be represented by extra speed neurons besides the CAN, where speed neurons' responses represent the moving speed (1D translation group parameter), and their feedback connections to the CAN represent the translation generator (Lie algebra). We demonstrated that the network responses are consistent with experimental data. Our model for the first time demonstrates how recurrent neural circuitry in the brain achieves equivariant stimulus representation.

## [On Gap-dependent Bounds for Offline Reinforcement Learning](#)

- Xinqi Wang · Qiwen Cui · Simon Du
- abstract@[open-review](#): This paper presents a systematic study on gap-dependent sample complexity in offline reinforcement learning. Prior works showed when the density ratio between an optimal policy and the behavior policy is upper bounded (single policy coverage), then the agent can achieve an  $\left(\frac{1}{\epsilon^2}\right)$  rate, which is also minimax optimal. We show under the same single policy coverage assumption, the rate can be improved to  $\left(\frac{1}{\epsilon}\right)$  when there is a gap in the optimal  $Q$ -function. Furthermore, we show under a stronger uniform single

policy coverage assumption, the sample complexity can be further improved to  $\$O(1)$ . Lastly, we also present nearly-matching lower bounds to complement our gap-dependent upper bounds.

## [High-dimensional limit theorems for SGD: Effective dynamics and critical scaling](#)

- Gerard Ben Arous · Reza Gheissari · Aukosh Jagannath
- abstract@[open-review](#): We study the scaling limits of stochastic gradient descent (SGD) with constant step-size in the high-dimensional regime. We prove limit theorems for the trajectories of summary statistics (i.e., finite-dimensional functions) of SGD as the dimension goes to infinity. Our approach allows one to choose the summary statistics that are tracked, the initialization, and the step-size. It yields both ballistic (ODE) and diffusive (SDE) limits, with the limit depending dramatically on the former choices. We find a critical scaling regime for the step-size below which this ``effective dynamics'' matches gradient flow for the population loss, but at which, a new correction term appears which changes the phase diagram. About the fixed points of this effective dynamics, the corresponding diffusive limits can be quite complex and even degenerate. We demonstrate our approach on popular examples including estimation for spiked matrix and tensor models and classification via two-layer networks for binary and XOR-type Gaussian mixture models. These examples exhibit surprising phenomena including multimodal timescales to convergence as well as convergence to sub-optimal solutions with probability bounded away from zero from random (e.g., Gaussian) initializations.

## [Decomposing NeRF for Editing via Feature Field Distillation](#)

- Sosuke Kobayashi · Eiichi Matsumoto · Vincent Sitzmann
- abstract@[open-review](#): Emerging neural radiance fields (NeRF) are a promising scene representation for computer graphics, enabling high-quality 3D reconstruction and novel view synthesis from image observations. However, editing a scene represented by a NeRF is challenging, as the underlying connectionist representations such as MLPs or voxel grids are not object-centric or compositional. In particular, it has been difficult to selectively edit specific regions or objects. In this work, we tackle the problem of semantic scene decomposition of NeRFs to enable query-based local editing of the represented 3D scenes. We propose to distill the knowledge of off-the-shelf, self-supervised 2D image feature extractors such as CLIP-LSeg or DINO into a 3D feature field optimized in parallel to the radiance field. Given a user-specified query of various modalities such as text, an image patch, or a point-and-click selection, 3D feature fields semantically decompose 3D space without the need for re-training, and enables us to semantically select and edit regions in the radiance field. Our experiments validate that the distilled feature fields can transfer recent progress in 2D vision and language foundation models to 3D scene representations, enabling convincing 3D segmentation and selective editing of emerging neural graphics representations.

## [Acceleration in Distributed Sparse Regression](#)

- Marie Maros · Gesualdo Scutari
- abstract@[open-review](#): We study acceleration for distributed sparse regression in  $\{\text{it high-dimensions}\}$ , which allows the parameter size to exceed and grow faster than the sample size. When applicable, existing distributed algorithms employing acceleration perform poorly in this setting, theoretically and numerically. We propose a new accelerated distributed algorithm suitable for high-dimensions. The method couples a suitable instance of accelerated Nesterov's proximal gradient with consensus and gradient-tracking mechanisms, aiming at estimating locally the gradient of the empirical loss while enforcing agreement on the local estimates. Under standard assumptions on the statistical model and tuning parameters, the proposed method is proved to globally converge at  $\{\text{it linear}\}$  rate to an estimate that is within the  $\{\text{it statistical precision}\}$  of the model. The iteration complexity scales as  $\$mathcal{O}(\sqrt{\kappa})$ , while the communications per iteration are at most  $\widetilde{\mathcal{O}}(\log m/(1-\rho))$ , where  $\kappa$  is the restricted condition number of the empirical loss,  $m$  is the number of agents, and  $\rho \in (0,1)$  measures the network connectivity. As by-product of our design, we also report an accelerated method for high-dimensional estimations over master-worker architectures, which is of independent interest and compares favorably with existing works.

## [Inference and Sampling for Archimax Copulas](#)

- Yuting Ng · Ali Hasan · Vahid Tarokh
- abstract@[open-review](#): Understanding multivariate dependencies in both the bulk and the tails of a distribution is an important problem for many applications, such as ensuring algorithms are robust to observations that are infrequent but have devastating effects. Archimax copulas are a family of distributions endowed with a precise representation that allows simultaneous modeling of the bulk and the tails of a distribution. Rather than separating the two as is typically done in practice, incorporating additional information from the bulk may improve inference of the tails, where observations are limited. Building on the stochastic representation of Archimax copulas, we develop a non-parametric inference method and sampling algorithm. Our proposed methods, to the best of our knowledge, are the first that allow for highly flexible and scalable inference and sampling algorithms, enabling the increased use of Archimax copulas in practical settings. We experimentally compare to state-of-the-art density modeling techniques, and the results suggest that the proposed method effectively extrapolates to tails while scaling to higher dimensional data. Our findings suggest that the proposed algorithms can be used in a variety of applications where understanding the interplay between the bulk and the tails of a distribution is necessary, such as health and safety.

## [Parameter tuning and model selection in Optimal Transport with semi-dual Brenier formulation](#)

- Adrien Vacher · Francois-Xavier Vialard
- abstract@[open-review](#): Over the past few years, numerous computational models have been developed to solve Optimal Transport (OT) in a stochastic setting, where distributions are represented by samples and where the goal is to find the closest map to the ground truth OT map, unknown in practical settings. So far, no quantitative criterion has yet been put forward to tune the parameter of these models and select maps that best approximate the ground truth. To perform this task, we propose to leverage the Brenier formulation of OT. Theoretically, we show that this formulation guarantees that, up to sharp a distortion parameter depending on the smoothness/strong convexity and a statistical deviation term, the selected map achieves the lowest quadratic error to the ground truth. This criterion, estimated via convex optimization, enables parameter tuning and model selection among entropic regularization of OT, input convex neural networks and smooth and strongly convex nearest-Brenier (SSNB) models. We also use this criterion to question the use of OT in Domain-Adaptation (DA). In a standard DA experiment, it enables us to identify the potential that is closest to the true OT map between the source and the target. Yet, we observe that this selected potential is far from being the one that performs best for the downstream transfer classification task.

## [Extracting computational mechanisms from neural data using low-rank RNNs](#)

- Adrian Valente · Jonathan Pillow · Srdjan Ostojic
- abstract@[open-review](#): An influential framework within systems neuroscience posits that neural computations can be understood in terms of low-dimensional dynamics in recurrent circuits. A number of methods have thus been developed to extract latent dynamical systems from neural recordings, but inferring models that are both predictive and interpretable remains a difficult challenge. Here we propose a new method called Low-rank Inference from Neural Trajectories (LINT), based on a class of low-rank recurrent neural networks (lrRNNs) for which a link between connectivity and dynamics has been previously demonstrated. By fitting such networks to trajectories of neural activity, LINT yields a mechanistic model of latent dynamics, as well as a set of axes for dimensionality reduction and verifiable predictions for inactivations of specific populations of neurons. Here, we first demonstrate the consistency of our method and apply it to two use cases: (i) we reverse-engineer "black-box" vanilla RNNs trained to perform cognitive tasks, and (ii) we infer latent dynamics and neural contributions from electrophysiological recordings of nonhuman primates performing a similar task.

## [CageNeRF: Cage-based Neural Radiance Field for Generalized 3D Deformation and Animation](#)

- Yicong Peng · Yichao Yan · Shengqi Liu · Yuhao Cheng · Shanyan Guan · Bowen Pan · Guangtao Zhai · Xiaokang Yang
- abstract@[open-review](#): While implicit representations have achieved high-fidelity results on 3D rendering, deforming and animating the learned implicit field remain a challenging task. Existing works typically leverage specific 3D model as deformation prior, such as SMPL for animating human. However, the category-specific prior dependency limits them to generalize to other objects. In this work, we propose a novel framework for deforming and animating the neural radiance field learned on arbitrary objects. The key insight is that we introduce a cage-based representation as deformation prior, which is category-agnostic. Specifically, the deformation is performed based on an enclosing cage with sparsely defined vertices inside the rendering space, where each point is projected into a novel position based on the barycentric interpolation of the deformed cage vertices via weight functions. In this way, we transform cage into a generalized constraint, which is able to deform and animate arbitrary target while preserving geometry details. Based on extensive experiments, we demonstrate the effectiveness of our framework in the task of geometry editing, object animation and deformation transfer.

## [Spending Thinking Time Wisely: Accelerating MCTS with Virtual Expansions](#)

- Weirui Ye · Pieter Abbeel · Yang Gao
- abstract@[open-review](#): One of the most important AI research questions is to trade off computation versus performance since ``perfect rationality'' exists in theory but is impossible to achieve in practice. Recently, Monte-Carlo tree search (MCTS) has attracted considerable attention due to the significant performance improvement in various challenging domains. However, the expensive time cost during search severely restricts its scope for applications. This paper proposes the Virtual MCTS (V-MCTS), a variant of MCTS that spends more search time on harder states and less search time on simpler states adaptively. We give theoretical bounds of the proposed method and evaluate the performance and computations on \$9 \times 9\$ Go board games and Atari games. Experiments show that our method can achieve comparable performances to the original search algorithm while requiring less than 50% search time on average. We believe that this approach is a viable alternative for tasks under limited time and resources.

## [Robust Models are less Over-Confident](#)

- Julia Grabinski · Paul Gavrikov · Janis Keuper · Margret Keuper
- abstract@[open-review](#): Regardless of the success of convolutional neural networks (CNNs) in many academic benchmarks of computer vision tasks, their application in real-world is still facing fundamental challenges, like the inherent lack of robustness as unveiled by adversarial attacks. These attacks target to manipulate the network's prediction by adding a small amount of noise onto the input. In turn, adversarial training (AT) aims to achieve robustness against such attacks by including adversarial samples in the training set. However, a general analysis of the reliability and model calibration of these robust models beyond adversarial robustness is still pending. In this paper, we analyze a variety of adversarially trained models that achieve high robust accuracies when facing state-of-the-art attacks and we show that AT has an interesting side-effect: it leads to models that are significantly less overconfident with their decisions even on clean data than non-robust models. Further, our analysis of robust models shows that not only AT but also the model's building blocks (activation functions and pooling) have a strong influence on the models' confidence.

## [Adversarial Attack on Attackers: Post-Process to Mitigate Black-Box Score-Based Query Attacks](#)

- Sizhe Chen · Zhehao Huang · Qinghua Tao · Yingwen Wu · Cihang Xie · Xiaolin Huang
- abstract@[open-review](#): The score-based query attacks (SQAs) pose practical threats to deep neural networks by crafting adversarial perturbations within dozens of queries, only using the model's output scores. Nonetheless, we note that if the loss trend of the outputs is slightly perturbed, SQAs could be easily misled and thereby become much less effective. Following this idea, we propose a novel defense, namely Adversarial Attack on Attackers (AAA), to confound SQAs towards incorrect attack directions by slightly modifying the output logits. In this way, (1) SQAs are prevented regardless of the model's worst-case robustness; (2) the original model predictions are hardly changed, i.e., no degradation on clean accuracy; (3) the calibration of confidence scores can be improved simultaneously. Extensive experiments are provided to verify the above advantages. For example, by setting \$\ell\_\infty=8/255\$ on CIFAR-10, our proposed AAA helps WideResNet-28 secure 80.59% accuracy under Square attack (2500 queries), while the best prior defense (i.e., adversarial training) only attains 67.44%. Since AAA attacks SQA's general greedy strategy, such advantages of AAA over 8 defenses can be consistently observed on 8 CIFAR-10/ImageNet models under 6 SQAs, using different attack targets, bounds, norms, losses, and strategies. Moreover, AAA calibrates better without hurting the accuracy. Our code is available at <https://github.com/Sizhe-Chen/AAA>.

## [Surface Coverage Optimization in Unknown Environments by Volumetric Integration](#)

- Antoine Guedon · Vincent Lepetit · Pascal Monasse
- abstract@[open-review](#): Next Best View computation (NBV) is a long-standing problem in robotics, and consists in identifying the next most informative sensor position(s) for reconstructing a 3D object or scene efficiently and accurately. Like most current methods, we consider NBV prediction from a depth sensor. Learning-based methods relying on a volumetric representation of the scene are suitable for path planning, but do not scale well with the size of the scene and have lower accuracy than methods using a surface-based representation. However, the latter constrain the camera to a small number of poses. To obtain the advantages of both representations, we show that we can maximize surface metrics by Monte Carlo integration over a volumetric representation. Our method scales to large scenes and handles free camera motion: It takes as input an arbitrarily large point cloud gathered by a depth sensor like Lidar systems as well as camera poses to predict NBV. We demonstrate our approach on a novel dataset made of large and complex 3D scenes.

## [Finding Differences Between Transformers and ConvNets Using Counterfactual Simulation Testing](#)

- Nataniel Ruiz · Cihang Xie · Sarah Bargal · Kate Saenko · Stan Sclaroff
- abstract@[open-review](#): Contemporary deep neural networks tend to be evaluated on static test sets. One shortcoming of this is the fact that these deep neural networks cannot be easily evaluated for robustness issues with respect to specific scene variations. For example, it is hard to study the robustness of these networks to variations of object scale, object pose, scene lighting and 3D occlusions. The main reason is that collecting real datasets with fine-grained naturalistic variations can be extremely time-consuming and expensive. In this work, we present Counterfactual Simulation Testing, a counterfactual framework that allows us to study the robustness of neural networks with respect to some of these naturalistic variations by building realistic synthetic scenes that allow us to ask counterfactual questions to the models, ultimately providing answers to questions such as "Would your classification still be correct if the object were viewed from the top?" or "Would your classification still be correct if the object would be partially occluded by another object?". Our method allows for a fair comparison of the robustness of recently released, state-of-the-art Convolutional Neural Networks and Vision Transformers, with respect to these naturalistic variations. We find evidence that ConvNext is more robust to pose and scale variations than Swin, that ConvNext generalizes better to our simulated domain and that Swin handles partial occlusion better than ConvNext. We also find that robustness for all networks improves with network scale and with data scale and variety.

## [High-dimensional Asymptotics of Feature Learning: How One Gradient Step Improves the Representation](#)

- Jimmy Ba · Murat Erdogdu · Taiji Suzuki · Zhichao Wang · Denny Wu · Greg Yang
- abstract@[open-review](#): We study the first gradient descent step on the first-layer parameters  $\mathbf{W}$  in a two-layer neural network:  $f(\mathbf{x}) = \frac{1}{\sqrt{N}} \mathbf{a}^\top \sigma(\mathbf{W}^\top \mathbf{x})$ , where  $\mathbf{W} \in \mathbb{R}^{d \times N}$ ,  $\mathbf{a} \in \mathbb{R}^N$  are randomly initialized, and the training objective is the empirical

MSE loss:  $\frac{1}{n} \sum_{i=1}^n (f(\mathbf{x}_i) - y_i)^2$ . In the proportional asymptotic limit where  $n, d, N \rightarrow \infty$  at the same rate, and an idealized student-teacher setting where the teacher  $f$  is a single-index model, we compute the prediction risk of ridge regression on the conjugate kernel after one gradient step on  $W$  with learning rate  $\eta$ . We consider two scalings of the first step learning rate  $\eta$ . For small  $\eta$ , we establish a Gaussian equivalence property for the trained feature map, and prove that the learned kernel improves upon the initial random features model, but cannot defeat the best linear model on the input. Whereas for sufficiently large  $\eta$ , we prove that for certain  $f$ , the same ridge estimator on trained features can go beyond this "linear regime" and outperform a wide range of (fixed) kernels. Our results demonstrate that even one gradient step can lead to a considerable advantage over random features, and highlight the role of learning rate scaling in the initial phase of training.

## [Finite-Time Analysis of Adaptive Temporal Difference Learning with Deep Neural Networks](#)

- Tao Sun · Dongsheng Li · Bao Wang
- abstract@[open-review](#): Temporal difference (TD) learning with function approximations (linear functions or neural networks) has achieved remarkable empirical success, giving impetus to the development of finite-time analysis. As an accelerated version of TD, the adaptive TD has been proposed and proved to enjoy finite-time convergence under the linear function approximation. Existing numerical results have demonstrated the superiority of adaptive algorithms to vanilla ones. Nevertheless, the performance guarantee of adaptive TD with neural network approximation remains widely unknown. This paper establishes the finite-time analysis for the adaptive TD with multi-layer ReLU network approximation whose samples are generated from a Markov decision process. Our established theory shows that if the width of the deep neural network is large enough, the adaptive TD using neural network approximation can find the (optimal) value function with high probabilities under the same iteration complexity as TD in general cases. Furthermore, we show that the adaptive TD using neural network approximation, with the same width and searching area, can achieve theoretical acceleration when the stochastic semi-gradients decay fast.

## [Multi-Granularity Cross-modal Alignment for Generalized Medical Visual Representation Learning](#)

- Fuying Wang · Yuyin Zhou · Shujun WANG · Varut Vardhanabhuti · Lequan Yu
- abstract@[open-review](#): Learning medical visual representations directly from paired radiology reports has become an emerging topic in representation learning. However, existing medical image-text joint learning methods are limited by instance or local supervision analysis, ignoring disease-level semantic correspondences. In this paper, we present a novel Multi-Granularity Cross-modal Alignment (MGCA) framework for generalized medical visual representation learning by harnessing the naturally exhibited semantic correspondences between medical image and radiology reports at three different levels, i.e., pathological region-level, instance-level, and disease-level. Specifically, we first incorporate the instance-wise alignment module by maximizing the agreement between image-report pairs. Further, for token-wise alignment, we introduce a bidirectional cross-attention strategy to explicitly learn the matching between fine-grained visual tokens and text tokens, followed by contrastive learning to align them. More important, to leverage the high-level inter-subject relationship semantic (e.g., disease) correspondences, we design a novel cross-modal disease-level alignment paradigm to enforce the cross-modal cluster assignment consistency. Extensive experimental results on seven downstream medical image datasets covering image classification, object detection, and semantic segmentation tasks demonstrate the stable and superior performance of our framework.

## [Bridging the Gap Between Vision Transformers and Convolutional Neural Networks on Small Datasets](#)

- Zhiying Lu · Hongtao Xie · Chuanbin Liu · Yongdong Zhang
- abstract@[open-review](#): There still remains extreme performance gap between Vision Transformers (ViTs) and Convolutional Neural Networks (CNNs) when training from scratch on small datasets, which is concluded to the lack of inductive bias. In this paper, we further consider this problem and point out two weakness of ViTs in inductive biases, that is, the spatial relevance and diverse channel representation. First, on spatial aspect, objects are locally compact and relevant, thus fine-grained feature needs to be extracted from a token and its neighbours. While the lack of data hinders ViTs to attend the spatial relevance. Second, on channel aspect, representation exhibits diversity on different channels. But the scarce data can not enable ViTs to learn strong enough representation for accurate recognition. To this end, we propose Dynamic Hybrid Vision Transformer (DHVT) as the solution to enhance the two inductive biases. On spatial aspect, we adopt a hybrid structure, in which convolution is integrated into patch embedding and multi-layer perceptron module, forcing the model to capture the token features as well as theirs neighbouring features. On channel aspect, we introduce a dynamic feature aggregation module in MLP and a brand new "head token" design in the multi-head self-attention module to help re-calibrate channel representation and make different channel group representation interacts with each other. The fusion of weak channel representation forms a strong enough representation for classification. With this design, we successfully eliminate the performance gap between CNNs and ViTs, and our DHVT achieves a series of state-of-the-art performance with a lightweight model, 85.68% on CIFAR-100 with 22.8M parameters, 82.3% on ImageNet-1K with 24.0M parameters. Code will be released if accepted.

## [Beyond spectral gap: the role of the topology in decentralized learning](#)

- Thijs Vogels · Hadrien Hendrikx · Martin Jaggi
- abstract@[open-review](#): In data-parallel optimization of machine learning models, workers collaborate to improve their estimates of the model: more accurate gradients allow them to use larger learning rates and optimize faster. We consider the setting in which all workers sample from the same dataset, and communicate over a sparse graph (decentralized). In this setting, current theory fails to capture important aspects of real-world behavior. First, the spectral gap<sup>TM</sup> of the communication graph is not predictive of its empirical performance in (deep) learning. Second, current theory does not explain that collaboration enables larger learning rates than training alone. In fact, it prescribes smaller learning rates, which further decrease as graphs become larger, failing to explain convergence in infinite graphs. This paper aims to paint an accurate picture of sparsely-connected distributed optimization when workers share the same data distribution. We quantify how the graph topology influences convergence in a quadratic toy problem and provide theoretical results for general smooth and (strongly) convex objectives. Our theory matches empirical observations in deep learning, and accurately describes the relative merits of different graph topologies.

## [Sharper Convergence Guarantees for Asynchronous SGD for Distributed and Federated Learning](#)

- Anastasiia Koloskova · Sebastian Stich · Martin Jaggi
- abstract@[open-review](#): We study the asynchronous stochastic gradient descent algorithm, for distributed training over  $n$  workers that might be heterogeneous. In this algorithm, workers compute stochastic gradients in parallel at their own pace and return them to the server without any synchronization. Existing convergence rates of this algorithm for non-convex smooth objectives depend on the maximum delay  $\tau_{\max}$  and reach an  $\epsilon$ -stationary point after  $O(\sigma^2 \epsilon^{-2} + \tau_{\max} \epsilon^{-1})$  iterations, where  $\sigma$  is the variance of stochastic gradients. In this work (i) we obtain a tighter convergence rate of  $O(\sqrt{\tau_{\max} \tau_{\text{avg}}} \epsilon^{-2})$  without any change in the algorithm where  $\tau_{\text{avg}}$  is the average delay, which can be significantly smaller than  $\tau_{\max}$ . We also provide (ii) a simple delay-adaptive learning rate scheme, under which asynchronous SGD achieves a convergence rate of  $O(\sqrt{\sigma^2 \epsilon^{-2} + \tau_{\text{avg}} \epsilon^{-1}})$ , and does not require any extra hyperparameter tuning nor extra communications. Our result allows to show for the first time that asynchronous SGD is always faster than mini-batch SGD. In addition, (iii) we consider the case of heterogeneous functions motivated by federated learning applications and improve the convergence rate by proving a weaker dependence on the maximum delay compared to prior works.

## [Implicit Warping for Animation with Image Sets](#)

- Arun Mallya · Ting-Chun Wang · Ming-Yu Liu
- abstract@[open-review](#): We present a new implicit warping framework for image animation using sets of source images through the transfer of motion of a driving video. A single cross-modal attention layer is used to find correspondences between the source images and the driving image, choose the most appropriate features from different source images, and warp the selected features. This is in contrast to the existing methods that use explicit flow-based warping, which is designed for animation using a single source and does not extend well to multiple sources. The pick-and-choose capability of our framework helps it achieve state-of-the-art results on multiple datasets for image animation using both single and multiple source images.

## [Indicators of Attack Failure: Debugging and Improving Optimization of Adversarial Examples](#)

- Maura Pintor · Luca Demetrio · Angelo Sotgiu · Ambra Demontis · Nicholas Carlini · Battista Biggio · Fabio Roli
- abstract@[open-review](#): Evaluating robustness of machine-learning models to adversarial examples is a challenging problem. Many defenses have been shown to provide a false sense of robustness by causing gradient-based attacks to fail, and they have been broken under more rigorous evaluations. Although guidelines and best practices have been suggested to improve current adversarial robustness evaluations, the lack of automatic testing and debugging tools makes it difficult to apply these recommendations in a systematic manner. In this work, we overcome these limitations by: (i) categorizing attack failures based on how they affect the optimization of gradient-based attacks, while also unveiling two novel failures affecting many popular attack implementations and past evaluations; (ii) proposing six novel \emph{indicators of failure}, to automatically detect the presence of such failures in the attack optimization process; and (iii) suggesting a systematic protocol to apply the corresponding fixes. Our extensive experimental analysis, involving more than 15 models in 3 distinct application domains, shows that our indicators of failure can be used to debug and improve current adversarial robustness evaluations, thereby providing a first concrete step towards automatizing and systematizing them.

## [Biologically Inspired Dynamic Thresholds for Spiking Neural Networks](#)

- Jianchuan Ding · Bo Dong · Felix Heide · Yufei Ding · Yunduo Zhou · Baocai Yin · Xin Yang
- abstract@[open-review](#): The dynamic membrane potential threshold, as one of the essential properties of a biological neuron, is a spontaneous regulation mechanism that maintains neuronal homeostasis, i.e. the constant overall spiking firing rate of a neuron. As such, the neuron firing rate is regulated by a dynamic spiking threshold, which has been extensively studied in biology. Existing work in the machine learning community does not employ biologically plausible spiking threshold schemes. This work aims at bridging this gap by introducing a novel bioinspired dynamic energy-temporal threshold (BDETT) scheme for spiking neural networks (SNNs). The proposed BDETT scheme mirrors two biologically plausible observations: a dynamic threshold has 1) a positive correlation with the average membrane potential and 2) a negative correlation with the preceding rate of depolarization. We validate the effectiveness of the proposed BDETT on robot obstacle avoidance and continuous control tasks under both normal conditions and various degraded conditions, including noisy observations, weights, and dynamic environments. We find that the BDETT outperforms existing static and heuristic threshold approaches by significant margins in all tested conditions, and we confirm that the proposed bioinspired dynamic threshold scheme offers biologically plausible homeostasis to SNNs in complex real-world tasks.

## [CUP: Critic-Guided Policy Reuse](#)

- Jin Zhang · Siyuan Li · Chongjie Zhang
- abstract@[open-review](#): The ability to reuse previous policies is an important aspect of human intelligence. To achieve efficient policy reuse, a Deep Reinforcement Learning (DRL) agent needs to decide when to reuse and which source policies to reuse. Previous methods solve this problem by introducing extra components to the underlying algorithm, such as hierarchical high-level policies over source policies, or estimations of source policies' value functions on the target task. However, training these components induces either optimization non-stationarity or heavy sampling cost, significantly impairing the effectiveness of transfer. To tackle this problem, we propose a novel policy reuse algorithm called Critic-gUided Policy reuse (CUP), which avoids training any extra components and efficiently reuses source policies. CUP utilizes the critic, a common component in actor-critic methods, to evaluate and choose source policies. At each state, CUP chooses the source policy that has the largest one-step improvement over the current target policy, and forms a guidance policy. The guidance policy is theoretically guaranteed to be a monotonic improvement over the current target policy. Then the target policy is regularized to imitate the guidance policy to perform efficient policy search. Empirical results demonstrate that CUP achieves efficient transfer and significantly outperforms baseline algorithms.

## [Concept Embedding Models](#)

- Mateo Espinosa Zarlenga · Pietro Barbiero · Gabriele Ciravegna · Giuseppe Marra · Francesco Giannini · Michelangelo Diligenti · Zohreh Shams · Frederic Precioso · Stefano Melacci · Adrian Weller · Pietro Lī · Mateja Jamnik
- abstract@[open-review](#): Deploying AI-powered systems requires trustworthy models supporting effective human interactions, going beyond raw prediction accuracy. Concept bottleneck models promote trustworthiness by conditioning classification tasks on an intermediate level of human-like concepts. This enables human interventions which can correct mispredicted concepts to improve the model's performance. However, existing concept bottleneck models are unable to find optimal compromises between high task accuracy, robust concept-based explanations, and effective interventions on concepts---particularly in real-world conditions where complete and accurate concept supervisions are scarce. To address this, we propose Concept Embedding Models, a novel family of concept bottleneck models which goes beyond the current accuracy-vs-interpretability trade-off by learning interpretable high-dimensional concept representations. Our experiments demonstrate that Concept Embedding Models (1) attain better or competitive task accuracy w.r.t. standard neural models without concepts, (2) provide concept representations capturing meaningful semantics including and beyond their ground truth labels, (3) support test-time concept interventions whose effect in test accuracy surpasses that in standard concept bottleneck models, and (4) scale to real-world conditions where complete concept supervisions are scarce.

## [ResT V2: Simpler, Faster and Stronger](#)

- Qinglong Zhang · Yu-Bin Yang
- abstract@[open-review](#): This paper proposes ResTv2, a simpler, faster, and stronger multi-scale vision Transformer for visual recognition. ResTv2 simplifies the EMSA structure in ResTv1 (i.e., eliminating the multi-head interaction part) and employs an upsample operation to reconstruct the lost medium- and high-frequency information caused by the downsampling operation. In addition, we explore different techniques for better applying ResTv2 backbones to downstream tasks. We find that although combining EMSAv2 and window attention can greatly reduce the theoretical matrix multiply FLOPs, it may significantly decrease the computation density, thus causing lower actual speed. We comprehensively validate ResTv2 on ImageNet classification, COCO detection, and ADE20K semantic segmentation. Experimental results show that the proposed ResTv2 can outperform the recently state-of-the-art backbones by a large margin, demonstrating the potential of ResTv2 as solid backbones. The code and models will be made publicly available at \url{https://github.com/wofmanaf/ResT}.

## [Neural Sheaf Diffusion: A Topological Perspective on Heterophily and Oversmoothing in GNNs](#)

- Cristian Bodnar · Francesco Di Giovanni · Benjamin Chamberlain · Pietro Lī · Michael Bronstein
- abstract@[open-review](#): Cellular sheaves equip graphs with a "geometrical" structure by assigning vector spaces and linear maps to nodes and edges. Graph Neural Networks (GNNs) implicitly assume a graph with a trivial underlying sheaf. This choice is reflected in the structure of the graph Laplacian operator, the properties of the associated diffusion equation, and the characteristics of the convolutional models that discretise this equation. In this paper, we use cellular sheaf theory to show that the underlying geometry of the graph is deeply linked with the performance of GNNs in heterophilic settings and

their oversmoothing behaviour. By considering a hierarchy of increasingly general sheaves, we study how the ability of the sheaf diffusion process to achieve linear separation of the classes in the infinite time limit expands. At the same time, we prove that when the sheaf is non-trivial, discretised parametric diffusion processes have greater control than GNNs over their asymptotic behaviour. On the practical side, we study how sheaves can be learned from data. The resulting sheaf diffusion models have many desirable properties that address the limitations of classical graph diffusion equations (and corresponding GNN models) and obtain state-of-the-art results in heterophilic settings. Overall, our work provides new connections between GNNs and algebraic topology and would be of interest to both fields.

## [On the Identifiability of Nonlinear ICA: Sparsity and Beyond](#)

- Yujia Zheng · Ignavier Ng · Kun Zhang
- abstract@[open-review](#): Nonlinear independent component analysis (ICA) aims to recover the underlying independent latent sources from their observable nonlinear mixtures. How to make the nonlinear ICA model identifiable up to certain trivial indeterminacies is a long-standing problem in unsupervised learning. Recent breakthroughs reformulate the standard independence assumption of sources as conditional independence given some auxiliary variables (e.g., class labels and/or domain/time indexes) as weak supervision or inductive bias. However, nonlinear ICA with unconditional priors cannot benefit from such developments. We explore an alternative path and consider only assumptions on the mixing process, such as Structural Sparsity or Independent Influences. We show that under specific instantiations of such constraints, the independent latent sources can be identified from their nonlinear mixtures up to a permutation and a component-wise transformation, thus achieving nontrivial identifiability of nonlinear ICA without auxiliary variables. We provide estimation methods and validate the theoretical results experimentally. The results on image data suggest that our conditions may hold in a number of practical data generating processes.

## [Truncated Matrix Power Iteration for Differentiable DAG Learning](#)

- Zhen Zhang · Ignavier Ng · Dong Gong · Yuhang Liu · Ehsan Abbasnejad · Mingming Gong · Kun Zhang · Javen Qinfeng Shi
- abstract@[open-review](#): Recovering underlying Directed Acyclic Graph structures (DAG) from observational data is highly challenging due to the combinatorial nature of the DAG-constrained optimization problem. Recently, DAG learning has been cast as a continuous optimization problem by characterizing the DAG constraint as a smooth equality one, generally based on polynomials over adjacency matrices. Existing methods place very small coefficients on high-order polynomial terms for stabilization, since they argue that large coefficients on the higher-order terms are harmful due to numeric exploding. On the contrary, we discover that large coefficients on higher-order terms are beneficial for DAG learning, when the spectral radiiuses of the adjacency matrices are small, and that larger coefficients for higher order terms can approximate the DAG constraints much better than the small counterparts. Based on this, we propose a novel DAG learning method with efficient truncated matrix power iteration to approximate geometric series based DAG constraints. Empirically, our DAG learning method outperforms the previous state-of-the-arts in various settings, often by a factor of 3 or more in terms of structural Hamming distance.

## [Contrastive Learning as Goal-Conditioned Reinforcement Learning](#)

- Benjamin Eysenbach · Tianjun Zhang · Sergey Levine · Russ Salakhutdinov
- abstract@[open-review](#): In reinforcement learning (RL), it is easier to solve a task if given a good representation. While deep RL should automatically acquire such good representations, prior work often finds that learning representations in an end-to-end fashion is unstable and instead equip RL algorithms with additional representation learning parts (e.g., auxiliary losses, data augmentation). How can we design RL algorithms that directly acquire good representations? In this paper, instead of adding representation learning parts to an existing RL algorithm, we show (contrastive) representation learning methods are already RL algorithms in their own right. To do this, we build upon prior work and apply contrastive representation learning to action-labeled trajectories, in such a way that the (inner product of) learned representations exactly corresponds to a goal-conditioned value function. We use this idea to reinterpret a prior RL method as performing contrastive learning, and then use the idea to propose a much simpler method that achieves similar performance. Across a range of goal-conditioned RL tasks, we demonstrate that contrastive RL methods achieve higher success rates than prior non-contrastive methods. We also show that contrastive RL outperforms prior methods on image-based tasks, without using data augmentation or auxiliary objectives

## [Robust Rent Division](#)

- Dominik Peters · Ariel Procaccia · David Zhu
- abstract@[open-review](#): In fair rent division, the problem is to assign rooms to roommates and fairly split the rent based on roommates' reported valuations for the rooms. Envy-free rent division is the most popular application on the fair division website Spliddit. The standard model assumes that agents can correctly report their valuations for each room. In practice, agents may be unsure about their valuations, for example because they have had only limited time to inspect the rooms. Our goal is to find a robust rent division that remains fair even if agent valuations are slightly different from the reported ones. We introduce the lexislack solution, which selects a rent division that remains envy-free for valuations within as large a radius as possible of the reported valuations. We also consider robustness notions for valuations that come from a probability distribution, and use results from learning theory to show how we can find rent divisions that (almost) maximize the probability of being envy-free, or that minimize the expected envy. We show that an almost optimal allocation can be identified based on polynomially many samples from the valuation distribution. Finding the best allocation given these samples is NP-hard, but in practice such an allocation can be found using integer linear programming.

## [Bayesian Persuasion for Algorithmic Recourse](#)

- Keegan Harris · Valerie Chen · Joon Kim · Ameet Talwalkar · Hoda Heidari · Steven Wu
- abstract@[open-review](#): When subjected to automated decision-making, decision subjects may strategically modify their observable features in ways they believe will maximize their chances of receiving a favorable decision. In many practical situations, the underlying assessment rule is deliberately kept secret to avoid gaming and maintain competitive advantage. The resulting opacity forces the decision subjects to rely on incomplete information when making strategic feature modifications. We capture such settings as a game of Bayesian persuasion, in which the decision maker offers a form of recourse to the decision subject by providing them with an action recommendation (or signal) to incentivize them to modify their features in desirable ways. We show that when using persuasion, the decision maker and decision subject are never worse off in expectation, while the decision maker can be significantly better off. While the decision maker's problem of finding the optimal Bayesian incentive compatible (BIC) signaling policy takes the form of optimization over infinitely many variables, we show that this optimization can be cast as a linear program over finitely-many regions of the space of possible assessment rules. While this reformulation simplifies the problem dramatically, solving the linear program requires reasoning about exponentially-many variables, even in relatively simple cases. Motivated by this observation, we provide a polynomial-time approximation scheme that recovers a near-optimal signaling policy. Finally, our numerical simulations on semi-synthetic data empirically demonstrate the benefits of using persuasion in the algorithmic recourse setting.

## [AdaFocal: Calibration-aware Adaptive Focal Loss](#)

- Arindam Ghosh · Thomas Schaaf · Matthew Gormley
- abstract@[open-review](#): Much recent work has been devoted to the problem of ensuring that a neural network's confidence scores match the true probability of being correct, i.e. the calibration problem. Of note, it was found that training with focal loss leads to better calibrated neural networks than cross-entropy, while achieving the same level of accuracy \cite{mukhoti2020}. This success stems from focal loss regularizing the entropy of the model's

prediction (controlled by the parameter  $\gamma$ ), thereby reining in the model's overconfidence. Further improvement is expected if  $\gamma$  is selected independently for each training sample; Sample-Dependent Focal Loss (FLSD-53) in [\cite{mukhoti2020}](#) does just that, but is based on heuristics that do not generalize well to all dataset-model pairs. In this paper, we propose a calibration-aware adaptive focal loss called AdaFocal that utilizes the calibration properties of focal (and inverse-focal loss) and adaptively modifies  $\gamma_t$  for different groups of samples based on (1)  $\gamma_{t-1}$  from the previous step (2) the magnitude of the model's under/over-confidence. We evaluate AdaFocal on various image recognition tasks and one NLP task, covering a variety of network architectures, to confirm the improvement in calibration while achieving similar levels of accuracy. Additionally, models trained with AdaFocal are shown to exhibit a significant boost in out-of-distribution detection capability.

## [Convolutional Neural Networks on Graphs with Chebyshev Approximation, Revisited](#)

- Mingguo He · Zhewei Wei · Ji-Rong Wen
- abstract@[open-review](#): Designing spectral convolutional networks is a challenging problem in graph learning. ChebNet, one of the early attempts, approximates the spectral graph convolutions using Chebyshev polynomials. GCN simplifies ChebNet by utilizing only the first two Chebyshev polynomials while still outperforming it on real-world datasets. GPR-GNN and BernNet demonstrate that the Monomial and Bernstein bases also outperform the Chebyshev basis in terms of learning the spectral graph convolutions. Such conclusions are counter-intuitive in the field of approximation theory, where it is established that the Chebyshev polynomial achieves the optimum convergent rate for approximating a function. In this paper, we revisit the problem of approximating the spectral graph convolutions with Chebyshev polynomials. We show that ChebNet's inferior performance is primarily due to illegal coefficients learnt by ChebNet approximating analytic filter functions, which leads to over-fitting. We then propose ChebNetII, a new GNN model based on Chebyshev interpolation, which enhances the original Chebyshev polynomial approximation while reducing the Runge phenomenon. We conducted an extensive experimental study to demonstrate that ChebNetII can learn arbitrary graph convolutions and achieve superior performance in both full- and semi-supervised node classification tasks. Most notably, we scale ChebNetII to a billion graph ogbn-papers100M, showing that spectral-based GNNs have superior performance.

## [Deep Bidirectional Language-Knowledge Graph Pretraining](#)

- Michihiro Yasunaga · Antoine Bosselut · Hongyu Ren · Xikun Zhang · Christopher D Manning · Percy Liang · Jure Leskovec
- abstract@[open-review](#): Pretraining a language model (LM) on text helps various downstream NLP tasks. Recent works show that a knowledge graph (KG) can complement text data, offering structured background knowledge and scaffold useful for reasoning. However, these works are not pretrained to learn deep fusion of the two modalities at scale, limiting the potential to acquire fully joint representations of text and KG. Here we propose DRAGON (Deep Bidirectional Language-Knowledge Graph Pretraining), a self-supervised approach to pretraining a deeply joint language-knowledge model from raw text and KG at scale. Specifically, our model takes pairs of text segments and relevant KG subgraphs as input and bidirectionally fuses information from both modalities. We pretrain this model by unifying two self-supervised reasoning objectives, masked language modeling and KG link prediction. DRAGON outperforms existing LMs and LM+KG models on diverse downstream tasks including question answering across general and biomedical domains, with +5% absolute gain on average across the board. In particular, DRAGON achieves notable performance on complex reasoning about language and knowledge (+10% on questions involving long context or multi-step reasoning) and low-resource QA (+8% on OBQA and RiddleSense), and new state-of-the-art results on various BioNLP tasks.

## [SoteriaFL: A Unified Framework for Private Federated Learning with Communication Compression](#)

- Zhize Li · Haoyu Zhao · Boyue Li · Yuejie Chi
- abstract@[open-review](#): To enable large-scale machine learning in bandwidth-hungry environments such as wireless networks, significant progress has been made recently in designing communication-efficient federated learning algorithms with the aid of communication compression. On the other end, privacy preserving, especially at the client level, is another important desideratum that has not been addressed simultaneously in the presence of advanced communication compression techniques yet. In this paper, we propose a unified framework that enhances the communication efficiency of private federated learning with communication compression. Exploiting both general compression operators and local differential privacy, we first examine a simple algorithm that applies compression directly to differentially-private stochastic gradient descent, and identify its limitations. We then propose a unified framework SoteriaFL for private federated learning, which accommodates a general family of local gradient estimators including popular stochastic variance-reduced gradient methods and the state-of-the-art shifted compression scheme. We provide a comprehensive characterization of its performance trade-offs in terms of privacy, utility, and communication complexity, where SoteriaFL is shown to achieve better communication complexity without sacrificing privacy nor utility than other private federated learning algorithms without communication compression.

## [Exploring Length Generalization in Large Language Models](#)

- Cem Anil · Yuhuai Wu · Anders Andreassen · Aitor Lewkowycz · Vedant Misra · Vinay Ramasesh · Ambrose Sloane · Guy Gur-Ari · Ethan Dyer · Behnam Neyshabur
- abstract@[open-review](#): The ability to extrapolate from short problem instances to longer ones is an important form of out-of-distribution generalization in reasoning tasks, and is crucial when learning from datasets where longer problem instances are rare. These include theorem proving, solving quantitative mathematics problems, and reading/summarizing novels. In this paper, we run careful empirical studies exploring the length generalization capabilities of transformer-based language models. We first establish that naively finetuning transformers on length generalization tasks shows significant generalization deficiencies independent of model scale. We then show that combining pretrained large language models' in-context learning abilities with scratchpad prompting (asking the model to output solution steps before producing an answer) results in a dramatic improvement in length generalization. We run careful failure analyses on each of the learning modalities and identify common sources of mistakes that highlight opportunities in equipping language models with the ability to generalize to longer problems.

## [Fast Vision Transformers with HiLo Attention](#)

- Zizheng Pan · Jianfei Cai · Bohan Zhuang
- abstract@[open-review](#): Vision Transformers (ViTs) have triggered the most recent and significant breakthroughs in computer vision. Their efficient designs are mostly guided by the indirect metric of computational complexity, i.e., FLOPs, which however has a clear gap with the direct metric such as throughput. Thus, we propose to use the direct speed evaluation on the target platform as the design principle for efficient ViTs. Particularly, we introduce LITv2, a simple and effective ViT which performs favourably against the existing state-of-the-art methods across a spectrum of different model sizes with faster speed. At the core of LITv2 is a novel self-attention mechanism, which we dub HiLo. HiLo is inspired by the insight that high frequencies in an image capture local fine details and low frequencies focus on global structures, whereas a multi-head self-attention layer neglects the characteristic of different frequencies. Therefore, we propose to disentangle the high/low frequency patterns in an attention layer by separating the heads into two groups, where one group encodes high frequencies via self-attention within each local window, and another group performs the attention to model the global relationship between the average-pooled low-frequency keys from each window and each query position in the input feature map. Benefit from the efficient design for both groups, we show that HiLo is superior to the existing attention mechanisms by comprehensively benchmarking on FLOPs, speed and memory consumption on GPUs. Powered by HiLo, LITv2 serves as a strong backbone for mainstream vision tasks including image classification, dense detection and segmentation.

## [Bringing Image Scene Structure to Video via Frame-Clip Consistency of Object Tokens](#)

- Elad Ben Avraham · Roei Herzig · Karttikeya Mangalam · Amir Bar · Anna Rohrbach · Leonid Karlinsky · Trevor Darrell · Amir Globerson
- abstract@[open-review](#): Recent action recognition models have achieved impressive results by integrating objects, their locations and interactions. However, obtaining dense structured annotations for each frame is tedious and time-consuming, making these methods expensive to train and less scalable. At the same time, if a small set of annotated images is available, either within or outside the domain of interest, how could we leverage these for a video downstream task? We propose a learning framework StructureViT (SViT for short), which demonstrates how utilizing the structure of a small number of images only available during training can improve a video model. SViT relies on two key insights. First, as both images and videos contain structured information, we enrich a transformer model with a set of object tokens that can be used across images and videos. Second, the scene representations of individual frames in video should ``align'' with those of still images. This is achieved via a Frame-Clip Consistency loss, which ensures the flow of structured information between images and videos. We explore a particular instantiation of scene structure, namely a Hand-Object Graph, consisting of hands and objects with their locations as nodes, and physical relations of contact/no-contact as edges. SViT shows strong performance improvements on multiple video understanding tasks and datasets.

## [AutoMTL: A Programming Framework for Automating Efficient Multi-Task Learning](#)

- Lijun Zhang · Xiao Liu · Hui Guan
- abstract@[open-review](#): Multi-task learning (MTL) jointly learns a set of tasks by sharing parameters among tasks. It is a promising approach for reducing storage costs while improving task accuracy for many computer vision tasks. The effective adoption of MTL faces two main challenges. The first challenge is to determine what parameters to share across tasks to optimize for both memory efficiency and task accuracy. The second challenge is to automatically apply MTL algorithms to an arbitrary CNN backbone without requiring time-consuming manual re-implementation and significant domain expertise. This paper addresses the challenges by developing the first programming framework AutoMTL that automates efficient MTL model development for vision tasks. AutoMTL takes as inputs an arbitrary backbone convolutional neural network (CNN) and a set of tasks to learn, and automatically produces a multi-task model that achieves high accuracy and small memory footprint simultaneously. Experiments on three popular MTL benchmarks (CityScapes, NYUv2, Tiny-Taskonomy) demonstrate the effectiveness of AutoMTL over state-of-the-art approaches as well as the generalizability of AutoMTL across CNNs. AutoMTL is open-sourced and available at <https://github.com/zhanglijun95/AutoMTL>.

## [Tight Mutual Information Estimation With Contrastive Fenchel-Legendre Optimization](#)

- Qing Guo · Junya Chen · Dong Wang · Yuwei Yang · Xinwei Deng · Jing Huang · Larry Carin · Chenyang Tao · Fan Li
- abstract@[open-review](#): Successful applications of InfoNCE (Information Noise-Contrastive Estimation) and its variants have popularized the use of contrastive variational mutual information (MI) estimators in machine learning . While featuring superior stability, these estimators crucially depend on costly large-batch training, and they sacrifice bound tightness for variance reduction. To overcome these limitations, we revisit the mathematics of popular variational MI bounds from the lens of unnormalized statistical modeling and convex optimization. Our investigation yields a new unified theoretical framework encompassing popular variational MI bounds, and leads to a novel, simple, and powerful contrastive MI estimator we name FLO. Theoretically, we show that the FLO estimator is tight, and it converges under stochastic gradient descent. Empirically, the proposed FLO estimator overcomes the limitations of its predecessors and learns more efficiently. The utility of FLO is verified using extensive benchmarks, and we further inspire the community with novel applications in meta-learning. Our presentation underscores the foundational importance of variational MI estimation in data-efficient learning.

## [Improved Convergence Rate of Stochastic Gradient Langevin Dynamics with Variance Reduction and its Application to Optimization](#)

- Yuri Kinoshita · Taiji Suzuki
- abstract@[open-review](#): The stochastic gradient Langevin Dynamics is one of the most fundamental algorithms to solve sampling problems and non-convex optimization appearing in several machine learning applications. Especially, its variance reduced versions have nowadays gained particular attention. In this paper, we study two variants of this kind, namely, the Stochastic Variance Reduced Gradient Langevin Dynamics and the Stochastic Recursive Gradient Langevin Dynamics. We prove their convergence to the objective distribution in terms of KL-divergence under the sole assumptions of smoothness and Log-Sobolev inequality which are weaker conditions than those used in prior works for these algorithms. With the batch size and the inner loop length set to  $\$sqrt{n}$ , the gradient complexity to achieve an  $\$epsilon$ -precision is  $\$tilde{O}((n+dn^{1/2})\epsilon^{-1})\gamma^2 L^2\alpha^{-2})$ , which is an improvement from any previous analyses. We also show some essential applications of our result to non-convex optimization.

## [Multivariate Time-Series Forecasting with Temporal Polynomial Graph Neural Networks](#)

- Yijing Liu · Qinxian Liu · Jian-Wei Zhang · Haozhe Feng · Zhongwei Wang · Zihan Zhou · Wei Chen
- abstract@[open-review](#): Modeling multivariate time series (MTS) is critical in modern intelligent systems. The accurate forecast of MTS data is still challenging due to the complicated latent variable correlation. Recent works apply the Graph Neural Networks (GNNs) to the task, with the basic idea of representing the correlation as a static graph. However, predicting with a static graph causes significant bias because the correlation is time-varying in the real-world MTS data. Besides, there is no gap analysis between the actual correlation and the learned one in their works to validate the effectiveness. This paper proposes a temporal polynomial graph neural network (TPGNN) for accurate MTS forecasting, which represents the dynamic variable correlation as a temporal matrix polynomial in two steps. First, we capture the overall correlation with a static matrix basis. Then, we use a set of time-varying coefficients and the matrix basis to construct a matrix polynomial for each time step. The constructed result empirically captures the precise dynamic correlation of six synthetic MTS datasets generated by a non-repeating random walk model. Moreover, the theoretical analysis shows that TPGNN can achieve perfect approximation under a commutative condition. We conduct extensive experiments on two traffic datasets with prior structure and four benchmark datasets. The results indicate that TPGNN achieves the state-of-the-art on both short-term and long-term MTS forecastings.

## [Feature-Proxy Transformer for Few-Shot Segmentation](#)

- Jian-Wei Zhang · Yifan Sun · Yi Yang · Wei Chen
- abstract@[open-review](#): Few-shot segmentation~(FSS) aims at performing semantic segmentation on novel classes given a few annotated support samples. With a rethink of recent advances, we find that the current FSS framework has deviated far from the supervised segmentation framework: Given the deep features, FSS methods typically use an intricate decoder to perform sophisticated pixel-wise matching, while the supervised segmentation methods use a simple linear classification head. Due to the intricacy of the decoder and its matching pipeline, it is not easy to follow such an FSS framework. This paper revives the straightforward framework of feature extractor `+$ linear classification head'` and proposes a novel Feature-Proxy Transformer (`FPTrans`) method, in which `the proxy` is the vector representing a semantic class in the linear classification head. FPTrans has two keypoints for learning discriminative features and representative proxies: 1) To better utilize the limited support samples, the feature extractor makes the query interact with the support features from bottom to top layers using a novel prompting strategy. 2) FPTrans uses multiple local background proxies (instead of a single one) because the background is not homogeneous and may contain some novel foreground regions. These two keypoints are easily integrated into the vision transformer backbone with the prompting mechanism in the transformer. Given the learned features and proxies, FPTrans directly compares their cosine similarity for segmentation. Although the framework is straightforward, we show that FPTrans achieves competitive FSS accuracy on par with state-of-the-art decoder-based methods.

## [Bayesian Risk Markov Decision Processes](#)

- Yifan Lin · Yuxuan Ren · Enlu Zhou
- abstract@[open-review](#): We consider finite-horizon Markov Decision Processes where parameters, such as transition probabilities, are unknown and estimated from data. The popular distributionally robust approach to addressing the parameter uncertainty can sometimes be overly conservative. In this paper, we propose a new formulation, Bayesian risk Markov decision process (BR-MDP), to address parameter uncertainty in MDPs, where a risk functional is applied in nested form to the expected total cost with respect to the Bayesian posterior distributions of the unknown parameters. The proposed formulation provides more flexible risk attitudes towards parameter uncertainty and takes into account the availability of data in future time stages. To solve the proposed formulation with the conditional value-at-risk (CVaR) risk functional, we propose an efficient approximation algorithm by deriving an analytical approximation of the value function and utilizing the convexity of CVaR. We demonstrate the empirical performance of the BR-MDP formulation and proposed algorithms on a gambler's betting problem and an inventory control problem.

## [Tsetlin Machine for Solving Contextual Bandit Problems](#)

- Raihan Seraj · Jivitesh Sharma · Ole-Christoffer Granmo
- abstract@[open-review](#): This paper introduces an interpretable contextual bandit algorithm using Tsetlin Machines, which solves complex pattern recognition tasks using propositional (Boolean) logic. The proposed bandit learning algorithm relies on straightforward bit manipulation, thus simplifying computation and interpretation. We then present a mechanism for performing Thompson sampling with Tsetlin Machine, given its non-parametric nature. Our empirical analysis shows that Tsetlin Machine as a base contextual bandit learner outperforms other popular base learners on eight out of nine datasets. We further analyze the interpretability of our learner, investigating how arms are selected based on propositional expressions that model the context.

## [Towards Scalable \(All-Pair\) Message Passing for Node Classification beyond Explicit Topology](#)

- Qitian Wu · Wentao Zhao · Zenan Li · David P Wipf · Junchi Yan
- abstract@[open-review](#): Graph neural networks have been extensively studied for learning with inter-connected data. Despite this, recent evidence has revealed GNNs' deficiencies related to over-squashing, heterophily, handling long-range dependencies, edge incompleteness and particularly, the absence of graphs altogether. While a plausible solution is to learn new topology for message passing, issues concerning quadratic complexity hinder simultaneous guarantees for scalability and precision in large networks. In this paper, we introduce a novel all-pair message passing scheme for efficiently propagating layer-wise signals between arbitrary nodes. Specifically, the efficient computation per layer is enabled by a kernelized Gumbel-Softmax operator that reduces the algorithmic complexity to linearity w.r.t. node numbers for learning latent structures from large, potentially fully-connected graphs in a differentiable manner. We also provide accompanying theory as justification for our design. Extensive experiments demonstrate the promising efficacy of the method in various tasks including node classification on different sizes of graphs (1K~1M) and graph-enhanced applications where input topology is missing.

## [Adaptive Data Debiasing through Bounded Exploration](#)

- Yifan Yang · Yang Liu · Parinaz Naghizadeh
- abstract@[open-review](#): Biases in existing datasets used to train algorithmic decision rules can raise ethical and economic concerns due to the resulting disparate treatment of different groups. We propose an algorithm for sequentially debiasing such datasets through adaptive and bounded exploration in a classification problem with costly and censored feedback. Exploration in this context means that at times, and to a judiciously-chosen extent, the decision maker deviates from its (current) loss-minimizing rule, and instead accepts some individuals that would otherwise be rejected, so as to reduce statistical data biases. Our proposed algorithm includes parameters that can be used to balance between the ultimate goal of removing data biases -- which will in turn lead to more accurate and fair decisions, and the exploration risks incurred to achieve this goal. We analytically show that such exploration can help debias data in certain distributions. We further investigate how fairness criteria can work in conjunction with our data debiasing algorithm. We illustrate the performance of our algorithm using experiments on synthetic and real-world datasets.

## [MSR: Making Self-supervised learning Robust to Aggressive Augmentations](#)

- Yingbin Bai · Erkun Yang · Zhaoqing Wang · Yuxuan Du · Bo Han · Cheng Deng · Dadong Wang · Tongliang Liu
- abstract@[open-review](#): Most recent self-supervised learning methods learn visual representation by contrasting different augmented views of images. Compared with supervised learning, more aggressive augmentations have been introduced to further improve the diversity of training pairs. However, aggressive augmentations may distort images' structures leading to a severe semantic shift problem that augmented views of the same image may not share the same semantics, thus degrading the transfer performance. To address this problem, we propose a new SSL paradigm, which counteracts the impact of semantic shift by balancing the role of weak and aggressively augmented pairs. Specifically, semantically inconsistent pairs are of minority and we treat them as noisy pairs. Note that deep neural networks (DNNs) have a crucial memorization effect that DNNs tend to first memorize clean (majority) examples before overfitting to noisy (minority) examples. Therefore, we set a relatively large weight for aggressively augmented data pairs at the early learning stage. With the training going on, the model begins to overfit noisy pairs. Accordingly, we gradually reduce the weights of aggressively augmented pairs. In doing so, our method can better embrace the aggressive augmentations and neutralize the semantic shift problem. Experiments show that our model achieves 73.1% top-1 accuracy on ImageNet-1K with ResNet-50 for 200 epochs, which is a 2.5% improvement over BYOL. Moreover, experiments also demonstrate that the learned representations can transfer well for various downstream tasks.

## [Exploiting Semantic Relations for Glass Surface Detection](#)

- Yuen-Hei Yeung · Jiaying Lin · Rynson Lau
- abstract@[open-review](#): Glass surfaces are omnipresent in our daily lives and often go unnoticed by the majority of us. While humans are generally able to infer their locations and thus avoid collisions, it can be difficult for current object detection systems to handle them due to the transparent nature of glass surfaces. Previous methods approached the problem by extracting global context information to obtain priors such as boundary and reflection. However, their performances cannot be guaranteed when these critical features are not available. We observe that humans often reason through the semantic context of the environment, which offers insights into the categories of and proximity between entities that are expected to appear in the surrounding. For example, the odds of co-occurrence of glass windows with walls and curtains is generally higher than that with other objects such as cars and trees, which have relatively less semantic relevance. Based on this observation, we propose a model that integrates the contextual relationship of the scene for glass surface detection with two novel modules: (1) Scene Aware Activation (SAA) Module to adaptively filter critical channels with respect to spatial and semantic features, and (2) Context Correlation Attention (CCA) Module to progressively learn the contextual correlations among objects both spatially and semantically. In addition, we propose a large-scale glass surface detection dataset named GSD-S, which contains 4,519 real-world RGB glass surface images from diverse real-world scenes with detailed annotations. Experimental results show that our model outperforms contemporary works, especially with 48.8% improvement on MAE from our proposed GSD-S dataset.

## [A Variant of Anderson Mixing with Minimal Memory Size](#)

- Fuchao Wei · Chenglong Bao · Yang Liu · Guangwen Yang
- abstract@[open-review](#): Anderson mixing (AM) is an acceleration method for fixed-point problems by exploring the information from historical iterations. Despite its numerical success in various applications, the memory requirement in AM remains a bottleneck when solving large-scale optimization problems in a resource-limited machine. To address this problem, we propose a novel variant of AM method, called Min-AM, by storing only one vector

pair, that is the minimal memory size requirement in AM. Our method forms a symmetric approximation to the inverse Hessian matrix and is proved to be equivalent to the full-memory Type-I AM for solving strongly convex quadratic optimization. Moreover, for general nonlinear optimization problems, we establish the convergence properties of Min-AM under reasonable assumptions and show that the mixing parameters can be adaptively chosen by estimating the eigenvalues of the Hessian. Finally, we extend Min-AM to solve stochastic programming problems. Experimental results on logistic regression and network training problems validate the effectiveness of the proposed Min-AM.

## [A Unified Framework for Deep Symbolic Regression](#)

- Mikel Landajuela · Chak Shing Lee · Jiachen Yang · Ruben Glatt · Claudio P Santiago · Ignacio Aravena · Terrell Mundhenk · Garrett Mulcahy · Brenden K Petersen
- abstract@[open-review](#): The last few years have witnessed a surge in methods for symbolic regression, from advances in traditional evolutionary methods to novel deep learning-based methods. Individual works typically focus on advancing the state-of-the-art for one particular class of solution methods, and there have been few attempts to investigate the benefits of hybridizing or integrating multiple methods. In this work, we identify five standalone symbolic regression methods whose individual capabilities---spanning neural-guided search, genetic programming, and linear regression---provide broad coverage of the overall space of existing approaches, and we propose a strategy to hybridize them into a single modular, unified symbolic regression framework. Based on empirical evaluation using SRBench, a new community tool for benchmarking symbolic regression methods, our unified framework achieves state-of-the-art performance in its ability to (1) symbolically recover analytical expressions, (2) fit datasets with high accuracy, and (3) balance accuracy-complexity trade-offs, across 252 ground-truth and black-box benchmark problems, in both noiseless settings and across various noise levels. Finally, we provide practical use case-based guidance for constructing hybrid symbolic regression algorithms, supported by extensive, combinatorial ablation studies.

## [Log-Polar Space Convolution Layers](#)

- Bing Su · Ji-Rong Wen
- abstract@[open-review](#): Convolutional neural networks use regular quadrilateral convolution kernels to extract features. Since the number of parameters increases quadratically with the size of the convolution kernel, many popular models use small convolution kernels, resulting in small local receptive fields in lower layers. This paper proposes a novel log-polar space convolution (LPSC) layer, where the convolution kernel is elliptical and adaptively divides its local receptive field into different regions according to the relative directions and logarithmic distances. The local receptive field grows exponentially with the number of distance levels. Therefore, the proposed LPSC not only naturally encodes local spatial structures, but also greatly increases the single-layer receptive field while maintaining the number of parameters. We show that LPSC can be implemented with conventional convolution via log-polar space pooling and can be applied in any network architecture to replace conventional convolutions. Experiments on different tasks and datasets demonstrate the effectiveness of the proposed LPSC.

## [MAtt: A Manifold Attention Network for EEG Decoding](#)

- Yue-Ting Pan · Jing-Lun Chou · Chun-Shu Wei
- abstract@[open-review](#): Recognition of electroencephalographic (EEG) signals highly affect the efficiency of non-invasive brain-computer interfaces (BCIs). While recent advances of deep-learning (DL)-based EEG decoders offer improved performances, the development of geometric learning (GL) has attracted much attention for offering exceptional robustness in decoding noisy EEG data. However, there is a lack of studies on the merged use of deep neural networks (DNNs) and geometric learning for EEG decoding. We herein propose a manifold attention network (mAtt), a novel geometric deep learning (GDL)-based model, featuring a manifold attention mechanism that characterizes spatiotemporal representations of EEG data fully on a Riemannian symmetric positive definite (SPD). The evaluation of the proposed mAtt on both time-synchronous and -asynchronous EEG datasets suggests its superiority over other leading DL methods for general EEG decoding. Furthermore, analysis of model interpretation reveals the capability of mAtt in capturing informative EEG features and handling the non-stationarity of brain dynamics.

## [Few-shot Image Generation via Adaptation-aware Kernel Modulation](#)

- Yunqing Zhao · Milad Abdollahzadeh · Keshigeyan Chandrasegaran · Ngai-Man (Man) Cheung
- abstract@[open-review](#): Few-shot image generation (FSIG) aims to learn to generate new and diverse samples given an extremely limited number of samples from a domain, e.g., 10 training samples. Recent work has addressed the problem using transfer learning approach, leveraging a GAN pretrained on a large-scale source domain dataset and adapting that model to the target domain based on very limited target domain samples. Central to recent FSIG methods are knowledge preserving criteria, which aim to select a subset of source model's knowledge to be preserved into the adapted model. However, a major limitation of existing methods is that their knowledge preserving criteria consider only source domain/source task, and they fail to consider target domain/adaptation task in selecting source model's knowledge, casting doubt on their suitability for setups of different proximity between source and target domain. Our work makes two contributions. As our first contribution, we re-visit recent FSIG works and their experiments. Our important finding is that, under setups which assumption of close proximity between source and target domains is relaxed, existing state-of-the-art (SOTA) methods which consider only source domain/source task in knowledge preserving perform no better than a baseline fine-tuning method. To address the limitation of existing methods, as our second contribution, we propose adaptation-aware kernel modulation to address general FSIG of different source-target domain proximity. Extensive experimental results show that the proposed method consistently achieves SOTA performance across source/target domains of different proximity, including challenging setups when source and target domains are more apart. Code / reproducibility details are included.

## [Discovering and Overcoming Limitations of Noise-engineered Data-free Knowledge Distillation](#)

- Piyush Vinod Raikwar · Deepak Mishra
- abstract@[open-review](#): Distillation in neural networks using only the samples randomly drawn from a Gaussian distribution is possibly the most straightforward solution one can think of for the complex problem of knowledge transfer from one network (teacher) to the other (student). If successfully done, it can eliminate the requirement of teacher's training data for knowledge distillation and avoid often arising privacy concerns in sensitive applications such as healthcare. There have been some recent attempts at Gaussian noise-based data-free knowledge distillation, however, none of them offer a consistent or reliable solution. We identify the shift in the distribution of hidden layer activation as the key limiting factor, which occurs when Gaussian noise is fed to the teacher network instead of the accustomed training data. We propose a simple solution to mitigate this shift and show that for vision tasks, such as classification, it is possible to achieve a performance close to the teacher by just using the samples randomly drawn from a Gaussian distribution. We validate our approach on CIFAR10, CIFAR100, SVHN, and Food101 datasets. We further show that in situations of sparsely available original data for distillation, the proposed Gaussian noise-based knowledge distillation method can outperform the distillation using the available data with a large margin. Our work lays the foundation for further research in the direction of noise-engineered knowledge distillation using random samples.

## [Theoretically Better and Numerically Faster Distributed Optimization with Smoothness-Aware Quantization Techniques](#)

- Bokun Wang · Mher Safaryan · Peter Richtarik
- abstract@[open-review](#): To address the high communication costs of distributed machine learning, a large body of work has been devoted in recent years to designing various compression strategies, such as sparsification and quantization, and optimization algorithms capable of using them. Recently, Safaryan et al. [2021] pioneered a dramatically different compression design approach: they first use the local training data to form local  $\{\mathbf{em}\}$  smoothness matrices} and then propose to design a compressor capable of exploiting the smoothness information contained therein. While this novel approach leads to substantial savings in communication, it is limited to sparsification as it crucially depends on the linearity of the compression operator. In this work, we

generalize their smoothness-aware compression strategy to {arbitrary unbiased compression} operators, which also includes sparsification. Specializing our results to stochastic quantization, we guarantee significant savings in communication complexity compared to standard quantization. In particular, we prove that block quantization with  $\mathcal{O}(n)$  blocks theoretically outperforms single block quantization, leading to a reduction in communication complexity by an  $\mathcal{O}(n)$  factor, where  $n$  is the number of nodes in the distributed system. Finally, we provide extensive numerical evidence with convex optimization problems that our smoothness-aware quantization strategies outperform existing quantization schemes as well as the aforementioned smoothness-aware sparsification strategies with respect to all relevant success measures: the number of iterations, the total amount of bits communicated, and wall-clock time.

## [Multiagent Q-learning with Sub-Team Coordination](#)

- Wenhan Huang · Kai Li · Kun Shao · Tianze Zhou · Matthew Taylor · Jun Luo · Dongge Wang · Hangyu Mao · Jianye Hao · Jun Wang · Xiaotie Deng
- abstract@[open-review](#): In many real-world cooperative multiagent reinforcement learning (MARL) tasks, teams of agents can rehearse together before deployment, but then communication constraints may force individual agents to execute independently when deployed. Centralized training and decentralized execution (CTDE) is increasingly popular in recent years, focusing mainly on this setting. In the value-based MARL branch, credit assignment mechanism is typically used to factorize the team reward into each individual's reward "individual-global-max (IGM) is a condition on the factorization ensuring that agents' action choices coincide with team's optimal joint action. However, current architectures fail to consider local coordination within sub-teams that should be exploited for more effective factorization, leading to faster learning. We propose a novel value factorization framework, called multiagent Q-learning with sub-team coordination (QSCAN), to flexibly represent sub-team coordination while honoring the IGM condition. QSCAN encompasses the full spectrum of sub-team coordination according to sub-team size, ranging from the monotonic value function class to the entire IGM function class, with familiar methods such as QMIX and QPLEX located at the respective extremes of the spectrum. Experimental results show that QSCAN's performance dominates state-of-the-art methods in matrix games, predator-prey tasks, the Switch challenge in MA-Gym. Additionally, QSCAN achieves comparable performances to those methods in a selection of StarCraft II micro-management tasks.

## [MEMO: Test Time Robustness via Adaptation and Augmentation](#)

- Marvin Zhang · Sergey Levine · Chelsea Finn
- abstract@[open-review](#): While deep neural networks can attain good accuracy on in-distribution test points, many applications require robustness even in the face of unexpected perturbations in the input, changes in the domain, or other sources of distribution shift. We study the problem of test time robustification, i.e., using the test input to improve model robustness. Recent prior works have proposed methods for test time adaptation, however, they each introduce additional assumptions, such as access to multiple test points, that prevent widespread adoption. In this work, we aim to study and devise methods that make no assumptions about the model training process and are broadly applicable at test time. We propose a simple approach that can be used in any test setting where the model is probabilistic and adaptable: when presented with a test example, perform different data augmentations on the data point, and then adapt (all of) the model parameters by minimizing the entropy of the model's average, or marginal, output distribution across the augmentations. Intuitively, this objective encourages the model to make the same prediction across different augmentations, thus enforcing the invariances encoded in these augmentations, while also maintaining confidence in its predictions. In our experiments, we evaluate two baseline ResNet models, two robust ResNet-50 models, and a robust vision transformer model, and we demonstrate that this approach achieves accuracy gains of 1-8% over standard model evaluation and also generally outperforms prior augmentation and adaptation strategies. For the setting in which only one test point is available, we achieve state-of-the-art results on the ImageNet-C, ImageNet-R, and, among ResNet-50 models, ImageNet-A distribution shift benchmarks.

## [Towards Debiased Learning and Out-of-Distribution Detection for Graph Data](#)

- Zenan Li · Qitian Wu · Fan Nie · Junchi Yan
- abstract@[open-review](#): Despite the remarkable success of graph neural networks (GNNs) for graph representation learning, they are generally built on the (unreliable) i.i.d. assumption across training and testing data. However, real-world graph data are universally comprised of outliers in training set and out-of-distribution (OOD) testing samples from unseen domains, which solicits effective models for i) debiased learning and ii) OOD detection, towards trustworthy general purpose. In this paper, we first mathematically formulate the two challenging problems for graph data and take an initiative on tackling them under a unified probabilistic model. Specifically, we model the graph generative process to characterize the distribution shifts of graph data together with an additionally introduced latent environment variable as an indicator. We then define a variational distribution, i.e., a recognition model, to infer the environment during training of GNN. By instantiating the generative models as two-component mixtures, we derive a tractable learning objective and theoretically justify that the model can i) automatically identify and down-weight outliers in the training procedure, and ii) induce an effective OOD detector from the recognition model. Experiments on diverse datasets with different types of OOD data prove that our model consistently outperforms strong baselines for both debiasing and OOD detection tasks. Our code will be made public when published.

## [Pruning has a disparate impact on model accuracy](#)

- Cuong Tran · Ferdinando Fioretto · Jung-Eun Kim · Rakshit Naidu
- abstract@[open-review](#): Network pruning is a widely-used compression technique that is able to significantly scale down overparameterized models with minimal loss of accuracy. This paper shows that pruning may create or exacerbate disparate impacts. The paper sheds light on the factors to cause such disparities, suggesting differences in gradient norms and distance to decision boundary across groups to be responsible for this critical issue. It analyzes these factors in detail, providing both theoretical and empirical support, and proposes a simple, yet effective, solution that mitigates the disparate impacts caused by pruning.

## [Structuring Uncertainty for Fine-Grained Sampling in Stochastic Segmentation Networks](#)

- Jakob Gawlikowski · Frank Nussbaum · Julia Niebling
- abstract@[open-review](#): In image segmentation, the classic approach of learning a deterministic segmentation neither accounts for noise and ambiguity in the data nor for expert disagreements about the correct segmentation. This has been addressed by architectures that predict heteroscedastic (input-dependent) segmentation uncertainty, which indicates regions of segmentations that should be treated with care. What is missing are structural insights into the uncertainty, which would be desirable for interpretability and systematic adjustments. In the context of state-of-the-art stochastic segmentation networks (SSNs), we solve this issue by dismantling the overall predicted uncertainty into smaller uncertainty components. We obtain them directly from the low-rank Gaussian distribution for the logits in the network head of SSNs, based on a previously unconsidered view of this distribution as a factor model. The rank subsequently encodes a number of latent variables, each of which controls an individual uncertainty component. Hence, we can use the latent variables (called factors) for fine-grained sample control, thereby solving an open problem from previous work. There is one caveat though--factors are only unique up to orthogonal rotations. Factor rotations allow us to structure the uncertainty in a way that endorses simplicity, non-redundancy, and separation among the individual uncertainty components. To make the overall and factor-specific uncertainties at play comprehensible, we introduce flow probabilities that quantify deviations from the mean prediction and can also be used for uncertainty visualization. We show on medical-imaging, earth-observation, and traffic-scene data that rotation criteria based on factor-specific flow probabilities consistently yield the best factors for fine-grained sampling.

## [Chroma-VAE: Mitigating Shortcut Learning with Generative Classifiers](#)

- Wanqian Yang · Polina Kirichenko · Micah Goldblum · Andrew Wilson
- abstract@[open-review](#): Deep neural networks are susceptible to shortcut learning, using simple features to achieve low training loss without discovering essential semantic structure. Contrary to prior belief, we show that generative models alone are not sufficient to prevent shortcut learning, despite an incentive to recover a more comprehensive representation of the data than discriminative approaches. However, we observe that shortcuts are preferentially encoded with minimal information, a fact that generative models can exploit to mitigate shortcut learning. In particular, we propose Chroma-VAE, a two-pronged approach where a VAE classifier is initially trained to isolate the shortcut in a small latent subspace, allowing a secondary classifier to be trained on the complementary, shortcut-free latent subspace. In addition to demonstrating the efficacy of Chroma-VAE on benchmark and real-world shortcut learning tasks, our work highlights the potential for manipulating the latent space of generative classifiers to isolate or interpret specific correlations.

## [Global Normalization for Streaming Speech Recognition in a Modular Framework](#)

- Ehsan Variani · Ke Wu · Michael D Riley · David Rybach · Matt Shannon · Cyril Allauzen
- abstract@[open-review](#): We introduce the Globally Normalized Autoregressive Transducer (GNAT) for addressing the label bias problem in streaming speech recognition. Our solution admits a tractable exact computation of the denominator for the sequence-level normalization. Through theoretical and empirical results, we demonstrate that by switching to a globally normalized model, the word error rate gap between streaming and non-streaming speech-recognition models can be greatly reduced (by more than 50% on the Librispeech dataset). This model is developed in a modular framework which encompasses all the common neural speech recognition models. The modularity of this framework enables controlled comparison of modelling choices and creation of new models.

## [Zero-Shot Video Question Answering via Frozen Bidirectional Language Models](#)

- Antoine Yang · Antoine Miech · Josef Sivic · Ivan Laptev · Cordelia Schmid
- abstract@[open-review](#): Video question answering (VideoQA) is a complex task that requires diverse multi-modal data for training. Manual annotation of question and answers for videos, however, is tedious and prohibits scalability. To tackle this problem, recent methods consider zero-shot settings with no manual annotation of visual question-answer. In particular, a promising approach adapts frozen autoregressive language models pretrained on Web-scale text-only data to multi-modal inputs. In contrast, we here build on frozen bidirectional language models (BiLM) and show that such an approach provides a stronger and cheaper alternative for zero-shot VideoQA. In particular, (i) we combine visual inputs with the frozen BiLM using light trainable modules, (ii) we train such modules using Web-scraped multi-modal data, and finally (iii) we perform zero-shot VideoQA inference through masked language modeling, where the masked text is the answer to a given question. Our proposed approach, FrozenBiLM, outperforms the state of the art in zero-shot VideoQA by a significant margin on a variety of datasets, including LSMDC-FiB, iVQA, MSRVTT-QA, MSVD-QA, ActivityNet-QA, TGIF-FrameQA, How2QA and TVQA. It also demonstrates competitive performance in the few-shot and fully-supervised setting. Our code and models will be made publicly available.

## [Singular Value Fine-tuning: Few-shot Segmentation requires Few-parameters Fine-tuning](#)

- Yanpeng Sun · Qiang Chen · Xiangyu He · Zechao Li · Jian Wang · Haocheng Feng · Junyu Han · Errui Ding · Jian Cheng · Jingdong Wang
- abstract@[open-review](#): Freezing the pre-trained backbone has become a standard paradigm to avoid overfitting in few-shot segmentation. In this paper, we rethink the paradigm and explore a new regime: {\\em fine-tuning a small part of parameters in the backbone}. We present a solution to overcome the overfitting problem, leading to better model generalization on learning novel classes. Our method decomposes backbone parameters into three successive matrices via the Singular Value Decomposition (SVD), then {\\em only fine-tunes the singular values} and keeps others frozen. The above design allows the model to adjust feature representations on novel classes while maintaining semantic clues within the pre-trained backbone. We evaluate our {\\em Singular Value Fine-tuning (SVF)} approach on various few-shot segmentation methods with different backbones. We achieve state-of-the-art results on both Pascal-5\$^i\$ and COCO-20\$^i\$ across 1-shot and 5-shot settings. Hopefully, this simple baseline will encourage researchers to rethink the role of backbone fine-tuning in few-shot settings.

## [Heatmap Distribution Matching for Human Pose Estimation](#)

- Haoxuan Qu · Li Xu · Yujun Cai · Lin Geng Foo · Jun Liu
- abstract@[open-review](#): For tackling the task of 2D human pose estimation, the great majority of the recent methods regard this task as a heatmap estimation problem, and optimize the heatmap prediction using the Gaussian-smoothed heatmap as the optimization objective and using the pixel-wise loss (e.g. MSE) as the loss function. In this paper, we show that optimizing the heatmap prediction in such a way, the model performance of body joint localization, which is the intrinsic objective of this task, may not be consistently improved during the optimization process of the heatmap prediction. To address this problem, from a novel perspective, we propose to formulate the optimization of the heatmap prediction as a distribution matching problem between the predicted heatmap and the dot annotation of the body joint directly. By doing so, our proposed method does not need to construct the Gaussian-smoothed heatmap and can achieve a more consistent model performance improvement during the optimization of the heatmap prediction. We show the effectiveness of our proposed method through extensive experiments on the COCO dataset and the MPII dataset.

## [High-dimensional Additive Gaussian Processes under Monotonicity Constraints](#)

- AndrÃ©s LÃ³pez-Lopera · Francois Bachoc · Olivier Roustant
- abstract@[open-review](#): We introduce an additive Gaussian process (GP) framework accounting for monotonicity constraints and scalable to high dimensions. Our contributions are threefold. First, we show that our framework enables to satisfy the constraints everywhere in the input space. We also show that more general componentwise linear inequality constraints can be handled similarly, such as componentwise convexity. Second, we propose the additive MaxMod algorithm for sequential dimension reduction. By sequentially maximizing a squared-norm criterion, MaxMod identifies the active input dimensions and refines the most important ones. This criterion can be computed explicitly at a linear cost. Finally, we provide open-source codes for our full framework. We demonstrate the performance and scalability of the methodology in several synthetic examples with hundreds of dimensions under monotonicity constraints as well as on a real-world flood application.

## [Bridging the Gap from Asymmetry Tricks to Decorrelation Principles in Non-contrastive Self-supervised Learning](#)

- Kang-Jun Liu · Masanori Suganuma · Takayuki Okatani
- abstract@[open-review](#): Recent non-contrastive methods for self-supervised representation learning show promising performance. While they are attractive since they do not need negative samples, it necessitates some mechanism to avoid collapsing into a trivial solution. Currently, there are two approaches to collapse prevention. One uses an asymmetric architecture on a joint embedding of input, e.g., BYOL and SimSiam, and the other imposes decorrelation criteria on the same joint embedding, e.g., Barlow-Twins and VICReg. The latter methods have theoretical support from information theory as to why they can learn good representation. However, it is not fully understood why the former performs equally well. In this paper, focusing on BYOL/SimSiam, which uses the stop-gradient and a predictor as asymmetric tricks, we present a novel interpretation of these tricks; they implicitly impose a constraint that encourages feature decorrelation similar to Barlow-Twins/VICReg. We then present a novel non-contrastive method, which replaces the stop-gradient in BYOL/SimSiam with the derived constraint; the method empirically shows comparable performance to the above SOTA methods in the standard

benchmark test using ImageNet. This result builds a bridge from BYOL/SimSiam to the decorrelation-based methods, contributing to demystifying their secrets.

## [Large-Scale Differentiable Causal Discovery of Factor Graphs](#)

- Romain Lopez · Jan-Christian Huetter · Jonathan Pritchard · Aviv Regev
- abstract@[open-review](#): A common theme in causal inference is learning causal relationships between observed variables, also known as causal discovery. This is usually a daunting task, given the large number of candidate causal graphs and the combinatorial nature of the search space. Perhaps for this reason, most research has so far focused on relatively small causal graphs, with up to hundreds of nodes. However, recent advances in fields like biology enable generating experimental data sets with thousands of interventions followed by rich profiling of thousands of variables, raising the opportunity and urgent need for large causal graph models. Here, we introduce the notion of factor directed acyclic graphs (\$f\$-DAGs) as a way to restrict the search space to non-linear low-rank causal interaction models. Combining this novel structural assumption with recent advances that bridge the gap between causal discovery and continuous optimization, we achieve causal discovery on thousands of variables. Additionally, as a model for the impact of statistical noise on this estimation procedure, we study edge perturbations of the \$f\$-DAG skeleton based on random graphs and quantify their effect on the \$f\$-DAG rank. This theoretical analysis suggests that the set of candidate \$f\$-DAGs is much smaller compared to the whole DAG space and thus more statistically robust in the high-dimensional regime where the underlying skeleton is hard to assess. We propose Differentiable Causal Discovery of Factor Graphs (DCD-FG), a scalable implementation of \$f\$-DAG constrained causal discovery for high-dimensional interventional data. DCD-FG uses a Gaussian non-linear low-rank structural equation model and shows significant improvements compared to state-of-the-art methods in both simulations as well as a recent large-scale single-cell RNA sequencing data set with hundreds of genetic interventions.

## [Learning Superpoint Graph Cut for 3D Instance Segmentation](#)

- Le Hui · Linghua Tang · Yaqi Shen · Jin Xie · Jian Yang
- abstract@[open-review](#): 3D instance segmentation is a challenging task due to complex local geometric structures of objects in point clouds. In this paper, we propose a learning-based superpoint graph cut method that explicitly learns the local geometric structures of the point cloud for instance segmentation. Specifically, we first oversegment the raw point clouds into superpoints and construct the superpoint graph. Then, we construct an edge score prediction network to predict the edge scores of the superpoint graph, where the similarity vectors of two adjacent nodes learned through cross-graph attention in the coordinate and feature spaces are used for regressing edge scores. By forcing two adjacent nodes of the same instance to be close to the instance center in the coordinate and feature spaces, we formulate a geometry-aware edge loss to train the edge score prediction network. Finally, we develop a superpoint graph cut network that employs the learned edge scores and the predicted semantic classes of nodes to generate instances, where bilateral graph attention is proposed to extract discriminative instance features on the coordinate and feature spaces for predicting semantic labels and scores of instances. Extensive experiments on two challenging datasets, ScanNet v2 and S3DIS, show that our method achieves new state-of-the-art performance.

## [Zero-Sum Stochastic Stackelberg Games](#)

- Denizalp Goktas · Sadie Zhao · Amy Greenwald
- abstract@[open-review](#): Min-max optimization problems (i.e., zero-sum games) have been used to model problems in a variety of fields in recent years, from machine learning to economics. The literature to date has mostly focused on static zero-sum games, assuming independent strategy sets. In this paper, we study a form of dynamic zero-sum games, called stochastic games, with dependent strategy sets. Just as zero-sum games with dependent strategy sets can be interpreted as zero-sum Stackelberg games, stochastic zero-sum games with dependent strategy sets can be interpreted as zero-sum stochastic Stackelberg games. We prove the existence of an optimal solution in zero-sum stochastic Stackelberg games (i.e., a recursive Stackelberg equilibrium), provide necessary and sufficient conditions for a solution to be optimal, and show that a recursive Stackelberg equilibrium can be computed in polynomial time via value iteration. Finally, we show that stochastic Stackelberg games can model the problem of pricing and allocating goods across agents and time; more specifically, we propose a stochastic Stackelberg game whose solutions correspond to a recursive competitive equilibrium in a stochastic Fisher market. We close with a series of experiments which confirm our theoretical results and show how value iteration performs in practice.

## [S-Prompts Learning with Pre-trained Transformers: An Occam's Razor for Domain Incremental Learning](#)

- Yabin Wang · Zhiwu Huang · Xiaopeng Hong
- abstract@[open-review](#): State-of-the-art deep neural networks are still struggling to address the catastrophic forgetting problem in continual learning. In this paper, we propose one simple paradigm (named as S-Prompting) and two concrete approaches to highly reduce the forgetting degree in one of the most typical continual learning scenarios, i.e., domain increment learning (DIL). The key idea of the paradigm is to learn prompts independently across domains with pre-trained transformers, avoiding the use of exemplars that commonly appear in conventional methods. This results in a win-win game where the prompting can achieve the best for each domain. The independent prompting across domains only requests one single cross-entropy loss for training and one simple K-NN operation as a domain identifier for inference. The learning paradigm derives an image prompt learning approach and a brand-new language-image prompt learning approach. Owning an excellent scalability (0.05% parameter increase per domain), the best of our approaches achieves a remarkable relative improvement (an average of about 30%) over the best of the state-of-the-art exemplar-free methods for three standard DIL tasks, and even surpasses the best of them relatively by about 6% in average when they use exemplars.

## [Random Sharpness-Aware Minimization](#)

- Yong Liu · Siqi Mai · Minhao Cheng · Xiangning Chen · Cho-Jui Hsieh · Yang You
- abstract@[open-review](#): Currently, Sharpness-Aware Minimization (SAM) is proposed to seek the parameters that lie in a flat region to improve the generalization when training neural networks. In particular, a minimax optimization objective is defined to find the maximum loss value centered on the weight, out of the purpose of simultaneously minimizing loss value and loss sharpness. For the sake of simplicity, SAM applies one-step gradient ascent to approximate the solution of the inner maximization. However, one-step gradient ascent may not be sufficient and multi-step gradient ascents will cause additional training costs. Based on this observation, we propose a novel random smoothing based SAM (R-SAM) algorithm. To be specific, R-SAM essentially smooths the loss landscape, based on which we are able to apply the one-step gradient ascent on the smoothed weights to improve the approximation of the inner maximization. Further, we evaluate our proposed R-SAM on CIFAR and ImageNet datasets. The experimental results illustrate that R-SAM can consistently improve the performance on ResNet and Vision Transformer (ViT) training.

## [Do Residual Neural Networks discretize Neural Ordinary Differential Equations?](#)

- Michael Sander · Pierre Ablin · Gabriel PeyrÃ©
- abstract@[open-review](#): Neural Ordinary Differential Equations (Neural ODEs) are the continuous analog of Residual Neural Networks (ResNets). We investigate whether the discrete dynamics defined by a ResNet are close to the continuous one of a Neural ODE. We first quantify the distance between the ResNet's hidden state trajectory and the solution of its corresponding Neural ODE. Our bound is tight and, on the negative side, does not go to \$0\$ with depth \$N\$ if the residual functions are not smooth with depth. On the positive side, we show that this smoothness is preserved by gradient descent for a ResNet with linear residual functions and small enough initial loss. It ensures an implicit regularization towards a limit Neural ODE at rate \$\frac{1}{N}\$, uniformly with depth and optimization time. As a byproduct of our analysis, we consider the use of a memory-free discrete adjoint method to train a ResNet by recovering the activations on the fly through a backward pass of the network, and show that this method theoretically succeeds at large depth if the residual functions are Lipschitz with the input. We then show that Heun's method, a second order ODE integration scheme, allows for better gradient

estimation with the adjoint method when the residual functions are smooth with depth. We experimentally validate that our adjoint method succeeds at large depth, and that Heunâ€™s method needs fewer layers to succeed. We finally use the adjoint method successfully for fine-tuning very deep ResNets without memory consumption in the residual layers.

## [A framework for bilevel optimization that enables stochastic and global variance reduction algorithms](#)

- Mathieu DagrÃ©ou · Pierre Ablin · Samuel Vaiter · Thomas Moreau
- abstract@[open-review](#): Bilevel optimization, the problem of minimizing a value function which involves the arg-minimum of another function, appears in many areas of machine learning. In a large scale empirical risk minimization setting where the number of samples is huge, it is crucial to develop stochastic methods, which only use a few samples at a time to progress. However, computing the gradient of the value function involves solving a linear system, which makes it difficult to derive unbiased stochastic estimates. To overcome this problem we introduce a novel framework, in which the solution of the inner problem, the solution of the linear system, and the main variable evolve at the same time. These directions are written as a sum, making it straightforward to derive unbiased estimates. The simplicity of our approach allows us to develop global variance reduction algorithms, where the dynamics of all variables is subject to variance reduction. We demonstrate that SABA, an adaptation of the celebrated SAGA algorithm in our framework, has  $\mathcal{O}(\frac{1}{T})$  convergence rate, and that it achieves linear convergence under Polyak-Lojasciewicz assumption. This is the first stochastic algorithm for bilevel optimization that verifies either of these properties. Numerical experiments validate the usefulness of our method.

## [Benchopt: Reproducible, efficient and collaborative optimization benchmarks](#)

- Thomas Moreau · Mathurin Massias · Alexandre Gramfort · Pierre Ablin · Pierre-Antoine Bannier · Benjamin Charlier · Mathieu DagrÃ©ou · Tom Dupre la Tour · Ghislain DURIF · Cassio F. Dantas · Quentin Klopfenstein · Johan Larsson · En Lai · Tanguy Lefort · BenoÃ®t MalÃ©zieux · Badr MOUFAD · Binh T. Nguyen · Alain Rakotomamonjy · Zaccharie Ramzi · Joseph Salmon · Samuel Vaiter
- abstract@[open-review](#): Numerical validation is at the core of machine learning research as it allows us to assess the actual impact of new methods, and to confirm the agreement between theory and practice. Yet, the rapid development of the field poses several challenges: researchers are confronted with a profusion of methods to compare, limited transparency and consensus on best practices, as well as tedious re-implementation work. As a result, validation is often very partial, which can lead to wrong conclusions that slow down the progress of research. We propose Benchopt, a collaborative framework to automatize, publish and reproduce optimization benchmarks in machine learning across programming languages and hardware architectures. Benchopt simplifies benchmarking for the community by providing an off-the-shelf tool for running, sharing and extending experiments. To demonstrate its broad usability, we showcase benchmarks on three standard ML tasks:  $\ell_2$ -regularized logistic regression, Lasso and ResNet18 training for image classification. These benchmarks highlight key practical findings that give a more nuanced view of state-of-the-art for these problems, showing that for practical evaluation, the devil is in the details.

## [A Non-asymptotic Analysis of Non-parametric Temporal-Difference Learning](#)

- EloÃ®se Berthier · Ziad Kobeissi · Francis Bach
- abstract@[open-review](#): Temporal-difference learning is a popular algorithm for policy evaluation. In this paper, we study the convergence of the regularized non-parametric TD(0) algorithm, in both the independent and Markovian observation settings. In particular, when TD is performed in a universal reproducing kernel Hilbert space (RKHS), we prove convergence of the averaged iterates to the optimal value function, even when it does not belong to the RKHS. We provide explicit convergence rates that depend on a source condition relating the regularity of the optimal value function to the RKHS. We illustrate this convergence numerically on a simple continuous-state Markov reward process.

## [Characterizing Datapoints via Second-Split Forgetting](#)

- Pratyush Maini · Saurabh Garg · Zachary Lipton · J. Zico Kolter
- abstract@[open-review](#): The dynamics by which neural networks learn and forget examples throughout training has emerged as an object of interest along several threads of research. In particular, researchers have proposed metrics of example hardness based on these dynamics, including (i) the epoch at which examples are first correctly classified; (ii) the number of times their predictions flip during training; and (iii) whether their prediction flips if they are held out. However, an example might be considered hard for several distinct reasons, such as being a member of a rare subpopulation, being mislabeled, or being fundamentally ambiguous in their class. In this paper, we focus on the second-split forgetting time (SSFT): the epoch (if any) after which an original training example is forgotten as the network is fine-tuned on a randomly held out partition of the data. Across multiple benchmark datasets and modalities, we demonstrate that mislabeled examples are forgotten quickly, and seemingly rare examples are forgotten comparatively slowly. By contrast, metrics only considering the first split learning dynamics struggle to differentiate the two. Additionally, the SSFT tends to be robust to the choice of architecture, optimizer, and random seed. From a practical standpoint, the SSFT (i) can help to identify mislabeled samples, the removal of which improves generalization; and (ii) can provide insights about failure modes. Through theoretical analysis addressing overparameterized linear models, we provide insights into how the observed phenomena may arise.

## [Tight Analysis of Extra-gradient and Optimistic Gradient Methods For Nonconvex Minimax Problems](#)

- Pouria Mahdavinia · Yuyang Deng · Haochuan Li · Mehrdad Mahdavi
- abstract@[open-review](#): Despite the established convergence theory of Optimistic Gradient Descent Ascent (OGDA) and Extragradient (EG) methods for the convex-concave minimax problems, little is known about the theoretical guarantees of these methods in nonconvex settings. To bridge this gap, for the first time, this paper establishes the convergence of OGDA and EG methods under the nonconvex-strongly-concave (NC-SC) and nonconvex-concave (NC-C) settings by providing a unified analysis through the lens of single-call extra-gradient methods. We further establish lower bounds on the convergence of GDA/OGDA/EG, shedding light on the tightness of our analysis. We also conduct experiments supporting our theoretical results. We believe our results will advance the theoretical understanding of OGDA and EG methods for solving complicated nonconvex minimax real-world problems, e.g., Generative Adversarial Networks (GANs) or robust neural networks training.

## [Identifying good directions to escape the NTK regime and efficiently learn low-degree plus sparse polynomials](#)

- Eshaan Nichani · Yu Bai · Jason Lee
- abstract@[open-review](#): A recent goal in the theory of deep learning is to identify how neural networks can escape the â€œlazy training,â€ or Neural Tangent Kernel (NTK) regime, where the network is coupled with its first order Taylor expansion at initialization. While the NTK is minimax optimal for learning dense polynomials (Ghorbani et al, 2021), it cannot learn features, and hence has poor sample complexity for learning many classes of functions including sparse polynomials. Recent works have thus aimed to identify settings where gradient based algorithms provably generalize better than the NTK. One such example is the â€œQuadNTKâ€ approach of Bai & Lee (2020), which analyzes the second-order term in the Taylor expansion. Bai & Lee (2020) show that the second-order term can learn sparse polynomials efficiently; however, it sacrifices the ability to learn general dense polynomials. In this paper, we analyze how gradient descent on a two-layer neural network can escape the NTK regime by utilizing a spectral characterization of the NTK (Montanari & Zhong, 2020) and building on the QuadNTK approach. We first expand upon the spectral analysis to identify â€œgoodâ€ directions in parameter space in which we can move without harming generalization. Next, we show that a wide two-layer neural network can jointly use the NTK and QuadNTK to fit target functions consisting of a dense low-degree term and a sparse high-degree term -- something neither the NTK nor the QuadNTK can do on their own. Finally, we construct a regularizer which encourages our parameter vector to move in the â€œgoodâ€

directions, and show that gradient descent on the regularized loss will converge to a global minimizer, which also has low test error. This yields an end-to-end convergence and generalization guarantee with provable sample complexity improvement over both the NTK and QuadNTK on their own.

## [RORL: Robust Offline Reinforcement Learning via Conservative Smoothing](#)

- Rui Yang · Chenjia Bai · Xiaoteng Ma · Zhaoran Wang · Chongjie Zhang · Lei Han
- abstract@[open-review](#): Offline reinforcement learning (RL) provides a promising direction to exploit the massive amount of offline data for complex decision-making tasks. Due to the distribution shift issue, current offline RL algorithms are generally designed to be conservative for value estimation and action selection. However, such conservatism impairs the robustness of learned policies, leading to a significant change even for a small perturbation on observations. To trade off robustness and conservatism, we propose Robust Offline Reinforcement Learning (RORL) with a novel conservative smoothing technique. In RORL, we explicitly introduce regularization on the policy and the value function for states near the dataset and additional conservative value estimation on these OOD states. Theoretically, we show RORL enjoys a tighter suboptimality bound than recent theoretical result in linear MDPs. We demonstrate that RORL can achieve the state-of-the-art performance on the general offline RL benchmark and is considerably robust to adversarial observation perturbation.

## [Optimal Scaling for Locally Balanced Proposals in Discrete Spaces](#)

- Haoran Sun · Hanjun Dai · Dale Schuurmans
- abstract@[open-review](#): Optimal scaling has been well studied for Metropolis-Hastings (M-H) algorithms in continuous spaces, but a similar understanding has been lacking in discrete spaces. Recently, a family of locally balanced proposals (LBP) for discrete spaces has been proved to be asymptotically optimal, but the question of optimal scaling has remained open. In this paper, we establish, for the first time, that the efficiency of M-H in discrete spaces can also be characterized by an asymptotic acceptance rate that is independent of the target distribution. Moreover, we verify, both theoretically and empirically, that the optimal acceptance rates for LBP and random walk Metropolis (RWM) are \$0.574\$ and \$0.234\$ respectively. These results also help establish that LBP is asymptotically \$O(N^{1/\frac{2}{3}})\$ more efficient than RWM with respect to model dimension \$N\$. Knowledge of the optimal acceptance rate allows one to automatically tune the neighborhood size of a proposal distribution in a discrete space, directly analogous to step-size control in continuous spaces. We demonstrate empirically that such adaptive M-H sampling can robustly improve sampling in a variety of target distributions in discrete spaces, including training deep energy based models.

## [Introspective Learning : A Two-Stage approach for Inference in Neural Networks](#)

- Mohit Prabhushankar · Ghassan AlRegib
- abstract@[open-review](#): In this paper, we advocate for two stages in a neural network's decision making process. The first is the existing feed-forward inference framework where patterns in given data are sensed and associated with previously learned patterns. The second stage is a slower reflection stage where we ask the network to reflect on its feed-forward decision by considering and evaluating all available choices. Together, we term the two stages as introspective learning. We use gradients of trained neural networks as a measurement of this reflection. A simple three-layered Multi Layer Perceptron is used as the second stage that predicts based on all extracted gradient features. We perceptually visualize the post-hoc explanations from both stages to provide a visual grounding to introspection. For the application of recognition, we show that an introspective network is \$4\%\$ more robust and \$42\%\$ less prone to calibration errors when generalizing to noisy data. We also illustrate the value of introspective networks in downstream tasks that require generalizability and calibration including active learning, out-of-distribution detection, and uncertainty estimation. Finally, we ground the proposed machine introspection to human introspection for the application of image quality assessment.

## [An Efficient Bayesian Data Augmentation Approach for Gradient-Bias Mitigation in Contrastive Learning](#)

- Changyou Chen · Jianyi Zhang · Yi Xu · Liqun Chen · Jiali Duan · Yiran Chen · Son Tran · Belinda Zeng · Trishul Chilimbi
- abstract@[open-review](#): Contrastive learning (CL) has been the de facto technique for self-supervised representation learning (SSL), with impressive empirical success such as multi-modal representation learning. However, traditional CL loss only considers negative samples from a minibatch, which could cause biased gradients due to the non-decomposability of the loss. For the first time, we consider optimizing a more generalized contrastive loss, where each data sample is associated with an infinite number of negative samples. We show that directly using minibatch stochastic optimization could lead to gradient bias. To remedy this, we propose an efficient Bayesian data augmentation technique to augment the contrastive loss into a decomposable one, where standard stochastic optimization can be directly applied without gradient bias. Specifically, our augmented loss defines a joint distribution over the model parameters and the augmented parameters, which can be conveniently optimized by a proposed stochastic expectation-maximization algorithm. Our framework is more general and is related to several popular SSL algorithms. We verify our framework on both small scale models and several large foundation models, including SSL of ImageNet and SSL for vision-language representation learning. Experiment results indicate the existence of gradient bias in all cases, and demonstrate the effectiveness of the proposed method on improving previous state of the arts. Remarkably, our method can outperform the strong MoCo-v3 under the same hyper-parameter setting, with only around half of the minibatch size.

## [Multimodal Contrastive Learning with LIMoE: the Language-Image Mixture of Experts](#)

- Basil Mustafa · Carlos Riquelme · Joan Puigcerver · Rodolphe Jenatton · Neil Houlsby
- abstract@[open-review](#): Large sparsely-activated models have obtained excellent performance in multiple domains. However, such models are typically trained on a single modality at a time. We present the Language-Image MoE, LIMoE, a sparse mixture of experts model capable of multimodal learning. LIMoE accepts both images and text simultaneously, while being trained using a contrastive loss. MoEs are a natural fit for a multimodal backbone, since expert layers can learn an appropriate partitioning of modalities. However, new challenges arise; in particular, training stability and balanced expert utilization, for which we propose an entropy-based regularization scheme. Across multiple scales, we demonstrate performance improvement over dense models of equivalent computational cost. LIMoE-L/16 trained comparably to CLIP-L/14 achieves 77.9% zero-shot ImageNet accuracy (vs. 76.2%), and when further scaled to H/14 (with additional data) it achieves 83.8%, approaching state-of-the-art methods which use custom per-modality backbones and pre-training schemes. We analyse the quantitative and qualitative behavior of LIMoE, and demonstrate phenomena such as differing treatment of the modalities and the emergence of modality-specific experts.

## [An empirical analysis of compute-optimal large language model training](#)

- Jordan Hoffmann · Sebastian Borgeaud · Arthur Mensch · Elena Buchatskaya · Trevor Cai · Eliza Rutherford · Diego de Las Casas · Lisa Anne Hendricks · Johannes Welbl · Aidan Clark · Thomas Hennigan · Eric Noland · Katherine Millican · George van den Driessche · Bogdan Damoc · Aurelia Guy · Simon Osindero · Karen Simonyan · Erich Elsen · Jack Rae · Oriol Vinyals · Laurent Sifre
- abstract@[open-review](#): We investigate the optimal model size and number of tokens for training a transformer language model under a given compute budget. We find that current large language models are significantly undertrained, a consequence of the recent focus on scaling language models whilst keeping the amount of training data constant. By training over 400 language models ranging from 70 million to over 16 billion parameters on 5 to 500 billion tokens, we find that for compute-optimal training, the model size and the number of training tokens should be scaled equally: for every doubling of model size the number of training tokens should also be doubled. We test this hypothesis by training a predicted compute-optimal model, Chinchilla, that uses the same compute budget as Gopher but with 70B parameters and 4\$times\$ more data. Chinchilla uniformly and significantly outperforms Gopher (280B), GPT-3 (175B), Jurassic-1 (178B), and Megatron-Turing NLG (530B) on a large range of downstream evaluation tasks. This also means that

Chinchilla uses substantially less compute for fine-tuning and inference, greatly facilitating downstream usage. As a highlight, Chinchilla reaches a state-of-the-art average accuracy of 67.5% on the MMLU benchmark, a 7% improvement over Gopher.

## [Temporal Latent Bottleneck: Synthesis of Fast and Slow Processing Mechanisms in Sequence Learning](#)

- Aniket Didolkar · Kshitij Gupta · Anirudh Goyal · Alex Lamb · Nan Rosemary Ke · Yoshua Bengio
- abstract@[open-review](#): Recurrent neural networks have a strong inductive bias towards learning temporally compressed representations, as the entire history of a sequence is represented by a single vector. By contrast, Transformers have little inductive bias towards learning temporally compressed representations, as they allow for attention over all previously computed elements in a sequence. Having a more compressed representation of a sequence may be beneficial for generalization, as a high-level representation may be more easily re-used and re-purposed and will contain fewer irrelevant details. At the same time, excessive compression of representations comes at the cost of expressiveness. We propose a solution which divides computation into two streams. A slow stream that is recurrent in nature aims to learn a specialized and compressed representation, by forcing chunks of \$K\$ time steps into a single representation which is divided into multiple vectors. At the same time, a fast stream is parameterized as a Transformer to process chunks consisting of \$K\$ time-steps conditioned on the information in the slow-stream. In the proposed approach we hope to gain the expressiveness of the Transformer, while encouraging better compression and structuring of representations in the slow stream. We show the benefits of the proposed method in terms of improved sample efficiency and generalization performance as compared to various competitive baselines for visual perception and sequential decision making tasks.

## [Data augmentation for efficient learning from parametric experts](#)

- Alexandre Galashov · Josh Merel · Nicolas Heess
- abstract@[open-review](#): We present a simple, yet powerful data-augmentation technique to enable data-efficient learning from parametric experts for reinforcement and imitation learning. We focus on what we call the policy cloning setting, in which we use online or offline queries of an expert or expert policy to inform the behavior of a student policy. This setting arises naturally in a number of problems, for instance as variants of behavior cloning, or as a component of other algorithms such as DAGGER, policy distillation or KL-regularized RL. Our approach, augmented policy cloning (APC), uses synthetic states to induce feedback-sensitivity in a region around sampled trajectories, thus dramatically reducing the environment interactions required for successful cloning of the expert. We achieve highly data-efficient transfer of behavior from an expert to a student policy for high-degrees-of-freedom control problems. We demonstrate the benefit of our method in the context of several existing and widely used algorithms that include policy cloning as a constituent part. Moreover, we highlight the benefits of our approach in two practically relevant settings (a) expert compression, i.e. transfer to a student with fewer parameters; and (b) transfer from privileged experts, i.e. where the expert has a different observation space than the student, usually including access to privileged information.

## [The Effects of Regularization and Data Augmentation are Class Dependent](#)

- Randall Balestrieri · Leon Bottou · Yann LeCun
- abstract@[open-review](#): Regularization is a fundamental technique to prevent over-fitting and to improve generalization performances by constraining a model's complexity. Current Deep Networks heavily rely on regularizers such as Data-Augmentation (DA) or weight-decay, and employ structural risk minimization, i.e. cross-validation, to select the optimal regularization hyper-parameters. In this study, we demonstrate that techniques such as DA or weight decay produce a model with a reduced complexity that is unfair across classes. The optimal amount of DA or weight decay found from cross-validation over all classes leads to disastrous model performances on some classes e.g. on Imagenet with a resnet50, the ``barn spider'' classification test accuracy falls from 68% to 46% only by introducing random crop DA during training. Even more surprising, such performance drop also appears when introducing uninformative regularization techniques such as weight decay. Those results demonstrate that our search for ever increasing generalization performance ---averaged over all classes and samples--- has left us with models and regularizers that silently sacrifice performances on some classes. This scenario can become dangerous when deploying a model on downstream tasks e.g. an Imagenet pre-trained resnet50 deployed on INaturalist sees its performances fall from 70% to 30% on class #8889 when introducing random crop DA during the Imagenet pre-training phase. Those results demonstrate that finding a correct measure of a model's complexity without class-dependent preference remains an open research question.

## [Meta-Learning Dynamics Forecasting Using Task Inference](#)

- Rui Wang · Robin Walters · Rose Yu
- abstract@[open-review](#): Current deep learning models for dynamics forecasting struggle with generalization. They can only forecast in a specific domain and fail when applied to systems with different parameters, external forces, or boundary conditions. We propose a model-based meta-learning method called DyAd which can generalize across heterogeneous domains by partitioning them into different tasks. DyAd has two parts: an encoder that infers the time-invariant hidden features of the task with weak supervision, and a forecaster which learns the shared dynamics of the entire domain. The encoder adapts and controls the forecaster during inference using adaptive instance normalization and adaptive padding. Theoretically, we prove that the generalization error of such a procedure is related to the task relatedness in the source domain, as well as the domain differences between source and target. Experimentally, we demonstrate that our model outperforms state-of-the-art approaches on forecasting complex physical dynamics including turbulent flow, real-world sea surface temperature, and ocean currents.

## [Jump Self-attention: Capturing High-order Statistics in Transformers](#)

- Haoyi Zhou · Siyang Xiao · Shanghang Zhang · Jieqi Peng · Shuai Zhang · Jianxin Li
- abstract@[open-review](#): The recent success of Transformer has benefited many real-world applications, with its capability of building long dependency through pairwise dot-products. However, the strong assumption that elements are directly attentive to each other limits the performance of tasks with high-order dependencies such as natural language understanding and Image captioning. To solve such problems, we are the first to define the Jump Self-attention (JAT) to build Transformers. Inspired by the pieces moving of English Draughts, we introduce the spectral convolutional technique to calculate JAT on the dot-product feature map. This technique allows JAT's propagation in each self-attention head and is interchangeable with the canonical self-attention. We further develop the higher-order variants under the multi-hop assumption to increase the generality. Moreover, the proposed architecture is compatible with the pre-trained models. With extensive experiments, we empirically show that our methods significantly increase the performance on ten different tasks.

## [ZeroC: A Neuro-Symbolic Model for Zero-shot Concept Recognition and Acquisition at Inference Time](#)

- Tailin Wu · Megan Tjandrasuwita · Zhengxuan Wu · Xuelin Yang · Kevin Liu · Rok Sosic · Jure Leskovec
- abstract@[open-review](#): Humans have the remarkable ability to recognize and acquire novel visual concepts in a zero-shot manner. Given a high-level, symbolic description of a novel concept in terms of previously learned visual concepts and their relations, humans can recognize novel concepts without seeing any examples. Moreover, they can acquire new concepts by parsing and communicating symbolic structures using learned visual concepts and relations. Endowing these capabilities in machines is pivotal in improving their generalization capability at inference time. In this work, we introduce Zero-shot Concept Recognition and Acquisition (ZeroC), a neuro-symbolic architecture that can recognize and acquire novel concepts in a zero-shot way. ZeroC represents concepts as graphs of constituent concept models (as nodes) and their relations (as edges). To allow inference time composition, we employ energy-based models (EBMs) to model concepts and relations. We design ZeroC architecture so that it allows a one-to-one mapping between a symbolic graph structure of a concept and its corresponding EBM, which allows acquiring new concepts, communicating its graph structure, and applying

it to classification and detection tasks at inference time. We introduce algorithms for learning and inference with ZeroC. We evaluate ZeroC on a challenging grid-world dataset which is designed to probe zero-shot concept recognition and acquisition, and demonstrate its capability.

## [Better Best of Both Worlds Bounds for Bandits with Switching Costs](#)

- Idan Amir · Guy Azov · Tomer Koren · Roi Livni
- abstract@[open-review](#): We study best-of-both-worlds algorithms for bandits with switching cost, recently addressed by Rouyer et al., 2021. We introduce a surprisingly simple and effective algorithm that simultaneously achieves minimax optimal regret bound of  $\mathcal{O}(T^{2/3})$  in the oblivious adversarial setting and a bound of  $\mathcal{O}(\min\{\log(T)\Delta^2, T^{2/3}\})$  in the stochastically-constrained regime, both with (unit) switching costs, where  $\Delta$  is the gap between the arms. In the stochastically constrained case, our bound improves over previous results due to Rouyer et al., 2021, that achieved regret of  $\mathcal{O}(T^{1/3}\Delta)$ . We accompany our results with a lower bound showing that, in general,  $\tilde{\mathcal{O}}(\Omega(\min\{1/\Delta^2, \Delta T\}))$  regret is unavoidable in the stochastically-constrained case.

## [You Never Stop Dancing: Non-freezing Dance Generation via Bank-constrained Manifold Projection](#)

- Jiangxin Sun · Chunyu Wang · Huang Hu · Hanjiang Lai · Zhi Jin · Jian-Fang Hu
- abstract@[open-review](#): One of the most overlooked challenges in dance generation is that the auto-regressive frameworks are prone to freezing motions due to noise accumulation. In this paper, we present two modules that can be plugged into the existing models to enable them to generate non-freezing and high fidelity dances. Since the high-dimensional motion data are easily swamped by noise, we propose to learn a low-dimensional manifold representation by an auto-encoder with a bank of latent codes, which can be used to reduce the noises in the predicted motions, thus preventing from freezing. We further extend the bank to provide explicit priors about the future motions to disambiguate motion prediction, which helps the predictors to generate motions with larger magnitude and higher fidelity than possible before. Extensive experiments on AIST++, a public large-scale 3D dance motion benchmark, demonstrate that our method notably outperforms the baselines in terms of quality, diversity and time length.

## [Vision Transformers learn patch association](#)

- Samy Jelassi · Michael Sander · Yuanzhi Li
- abstract@[open-review](#): Vision Transformers (ViTs) have recently achieved comparable or superior performance to Convolutional neural networks (CNNs) in computer vision. This empirical breakthrough is even more remarkable since ViTs discards spatial information by mixing patch embeddings and positional encodings and do not embed any visual inductive bias (e.g.\ spatial locality). Yet, recent work showed that while minimizing their training loss, ViTs specifically learn spatially delocalized patterns. This raises a central question: how do ViTs learn this pattern by solely minimizing their training loss using gradient-based methods from \emph{random initialization}? We propose a structured classification dataset and a simplified ViT model to provide preliminary theoretical justification of this phenomenon. Our model relies on a simplified attention mechanism --the positional attention mechanism-- where the attention matrix solely depends on the positional encodings. While the problem admits multiple solutions that generalize, we show that our model implicitly learns the spatial structure of the dataset while generalizing. We finally prove that learning the structure helps to sample-efficiently transfer to downstream datasets that share the same structure as the pre-training one but with different features. We empirically verify that ViTs using only the positional attention mechanism perform similarly to the original one on CIFAR-10/100, SVHN and ImageNet.

## [Tractable Function-Space Variational Inference in Bayesian Neural Networks](#)

- Tim G. J. Rudner · Zonghao Chen · Yee Whye Teh · Yarin Gal
- abstract@[open-review](#): Reliable predictive uncertainty estimation plays an important role in allowing neural networks to be deployed in safety-critical settings. A popular approach for estimating the predictive uncertainty of neural networks is to treat the network parameters as random variables and infer an approximate posterior that can be used to obtain a distribution over network predictions. However, explicit inference over neural network parameters makes it difficult to incorporate meaningful prior information about the data generating process into training. In this paper, we pursue an alternative approach. Noting that stochastic neural networks define distributions over functions induced by distributions over parameters, we follow prior work in framing Bayesian inference in neural networks as inferring a posterior distribution over functions and propose a scalable function-space variational inference method that allows incorporating prior information and encourages reliable predictive uncertainty estimation. We show that the proposed method leads to state-of-the-art uncertainty estimation and predictive performance on a range of prediction problems, and demonstrate that it performs well on a challenging safety-critical medical diagnosis task in which reliable uncertainty estimation is essential.

## [When are Offline Two-Player Zero-Sum Markov Games Solvable?](#)

- Qiwen Cui · Simon Du
- abstract@[open-review](#): We study what dataset assumption permits solving offline two-player zero-sum Markov games. In stark contrast to the offline single-agent Markov decision process, we show that the single strategy concentration assumption is insufficient for learning the Nash equilibrium (NE) strategy in offline two-player zero-sum Markov games. On the other hand, we propose a new assumption named unilateral concentration and design a pessimism-type algorithm that is provably efficient under this assumption. In addition, we show that the unilateral concentration assumption is necessary for learning an NE strategy. Furthermore, our algorithm can achieve minimax sample complexity without any modification for two widely studied settings: dataset with uniform concentration assumption and turn-based Markov games. Our work serves as an important initial step towards understanding offline multi-agent reinforcement learning.

## [Online Algorithms for the Santa Claus Problem](#)

- Max Springer · MohammadTaghi Hajiaghayi · Debmalya Panigrahi · Mohammad Khani
- abstract@[open-review](#): The Santa Claus problem is a fundamental problem in \emph{fair division}: the goal is to partition a set of \emph{heterogeneous} items among \emph{agents} so as to maximize the minimum value of items received by any agent. In this paper, we study the online version of this problem where the items are not known in advance and have to be assigned to agents as they arrive over time. If the arrival order of items is arbitrary, then no good assignment exists in the worst case. However, we show that even for arbitrary items, if their arrival order is random, then for any  $\epsilon > 0$ , we can obtain a competitive ratio of  $1-\epsilon$  when the optimal assignment gives value at least  $\Omega(\log n / \epsilon^2)$  to every agent. We also show that this result is almost tight: namely, if the optimal solution has value at most  $C \ln n / \epsilon$  for some constant  $C$ , then there is no  $(1-\epsilon)$ -competitive algorithm even with random arrival order.

## [Surprising Instabilities in Training Deep Networks and a Theoretical Analysis](#)

- Yuxin Sun · DONG LAO · Ganesh Sundaramoorthi · Anthony Yezzi
- abstract@[open-review](#): We empirically demonstrate numerical instabilities in training standard deep networks with SGD. Specifically, we show numerical error (on the order of the smallest floating point bit) induced from floating point arithmetic in training deep nets can be amplified significantly and result in significant test accuracy variance, comparable to the test accuracy variance due to stochasticity in SGD. We show how this is likely traced to instabilities of the optimization dynamics that are localized over iterations and regions of the weight tensor space. We do this by presenting a theoretical framework using numerical analysis of partial differential equations (PDE), and analyzing the gradient descent PDE of a one-layer convolutional neural network,

which is sufficient to illustrate these instabilities. We show that it is stable only under certain conditions on the learning rate and weight decay. We reproduce the localized instabilities in the PDE for the one-layer network, which arise when the conditions are violated.

## [Learning to Reason with Neural Networks: Generalization, Unseen Data and Boolean Measures](#)

- Emmanuel Abbe · Samy Bengio · Elisabetta Cornacchia · Jon Kleinberg · Aryo Lotfi · Maithra Raghu · Chiyuan Zhang
- abstract@[open-review](#): This paper considers the Pointer Value Retrieval (PVR) benchmark introduced in [ZRKB21], where a ‘reasoning’ function acts on a string of digits to produce the label. More generally, the paper considers the learning of logical functions with gradient descent (GD) on neural networks. It is first shown that in order to learn logical functions with gradient descent on symmetric neural networks, the generalization error can be lower-bounded in terms of the noise-stability of the target function, supporting a conjecture made in [ZRKB21]. It is then shown that in the distribution shift setting, when the data withholding corresponds to freezing a single feature (referred to as canonical holdout), the generalization error of gradient descent admits a tight characterization in terms of the Boolean influence for several relevant architectures. This is shown on linear models and supported experimentally on other models such as MLPs and Transformers. In particular, this puts forward the hypothesis that for such architectures and for learning logical functions such as PVR functions, GD tends to have an implicit bias towards low-degree representations, which in turn gives the Boolean influence for the generalization error under quadratic loss.

## [Hardness of Noise-Free Learning for Two-Hidden-Layer Neural Networks](#)

- Sitan Chen · Aravind Gollakota · Adam Klivans · Raghu Meka
- abstract@[open-review](#): We give superpolynomial statistical query (SQ) lower bounds for learning two-hidden-layer ReLU networks with respect to Gaussian inputs in the standard (noise-free) model. No general SQ lower bounds were known for learning ReLU networks of any depth in this setting: previous SQ lower bounds held only for adversarial noise models (agnostic learning) (Kothari and Klivans 2014, Goel et al. 2020a, Diakonikolas et al. 2020a) or restricted models such as correlational SQ (Goel et al. 2020b, Diakonikolas et al. 2020b). Prior work hinted at the impossibility of our result: Vempala and Wilmes (2019) showed that general SQ lower bounds cannot apply to any real-valued family of functions that satisfies a simple non-degeneracy condition. To circumvent their result, we refine a lifting procedure due to Daniely and Vardi (2021) that reduces Boolean PAC learning problems to Gaussian ones. We show how to extend their technique to other learning models and, in many well-studied cases, obtain a more efficient reduction. As such, we also prove new cryptographic hardness results for PAC learning two-hidden-layer ReLU networks, as well as new lower bounds for learning constant-depth ReLU networks from membership queries.

## [On the Tradeoff Between Robustness and Fairness](#)

- Xinsong Ma · Zekai Wang · Weiwei Liu
- abstract@[open-review](#): Interestingly, recent experimental results [2, 26, 22] have identified a robust fairness phenomenon in adversarial training (AT), namely that a robust model well-trained by AT exhibits a remarkable disparity of standard accuracy and robust accuracy among different classes compared with natural training. However, the effect of different perturbation radii in AT on robust fairness has not been studied, and one natural question is raised: does a tradeoff exist between average robustness and robust fairness? Our extensive experimental results provide an affirmative answer to this question: with an increasing perturbation radius, stronger AT will lead to a larger class-wise disparity of robust accuracy. Theoretically, we analyze the class-wise performance of adversarially trained linear models with mixture Gaussian distribution. Our theoretical results support our observations. Moreover, our theory shows that adversarial training easily leads to more serious robust fairness issue than natural training. Motivated by theoretical results, we propose a fairly adversarial training (FAT) method to mitigate the tradeoff between average robustness and robust fairness. Experimental results validate the effectiveness of our proposed method.

## [On Image Segmentation With Noisy Labels: Characterization and Volume Properties of the Optimal Solutions to Accuracy and Dice](#)

- Marcus Nordstrom · Henrik Hult · Fredrik Ljungman · Jonas Sjöderberg
- abstract@[open-review](#): We study two of the most popular performance metrics in medical image segmentation, Accuracy and Dice, when the target labels are noisy. For both metrics, several statements related to characterization and volume properties of the set of optimal segmentations are proved, and associated experiments are provided. Our main insights are: (i) the volume of the solutions to both metrics may deviate significantly from the expected volume of the target, (ii) the volume of a solution to Accuracy is always less than or equal to the volume of a solution to Dice and (iii) the optimal solutions to both of these metrics coincide when the set of feasible segmentations is constrained to the set of segmentations with the volume equal to the expected volume of the target.

## [WebShop: Towards Scalable Real-World Web Interaction with Grounded Language Agents](#)

- Shunyu Yao · Howard Chen · John Yang · Karthik Narasimhan
- abstract@[open-review](#): Most existing benchmarks for grounding language in interactive environments either lack realistic linguistic elements, or prove difficult to scale up due to substantial human involvement in the collection of data or feedback signals. We develop WebShop – a simulated e-commerce website environment with 1.18 million real-world products and 12,087 crowd-sourced text instructions. In this environment, an agent needs to navigate multiple types of webpages and issue diverse actions to find, customize, and purchase a product given an instruction. WebShop provides several challenges including understanding compositional instructions, query (re-)formulation, dealing with noisy text in webpages, and performing strategic exploration. We collect over 1,600 human trajectories to first validate the benchmark, then train and evaluate a diverse range of agents using reinforcement learning, imitation learning, and pre-trained image and language models. Our best model achieves a task success rate of 29%, which significantly outperforms rule heuristics but is far lower than expert human performance (59%). We also analyze agent and human trajectories and ablate various model components to provide insights for developing future agents with stronger language understanding and decision making abilities. Finally, we show our agent trained on WebShop exhibits non-trivial sim-to-real transfer when evaluated on amazon.com and ebay.com, indicating the potential value of our benchmark for developing practical web agents that can operate in the wild.

## [Q-ViT: Accurate and Fully Quantized Low-bit Vision Transformer](#)

- Yanjing Li · Sheng Xu · Baochang Zhang · Xianbin Cao · Peng Gao · Guodong Guo
- abstract@[open-review](#): The large pre-trained vision transformers (ViTs) have demonstrated remarkable performance on various visual tasks, but suffer from expensive computational and memory cost problems when deployed on resource-constrained devices. Among the powerful compression approaches, quantization extremely reduces the computation and memory consumption by low-bit parameters and bit-wise operations. However, low-bit ViTs remain largely unexplored and usually suffer from a significant performance drop compared with the real-valued counterparts. In this work, through extensive empirical analysis, we first identify the bottleneck for severe performance drop comes from the information distortion of the low-bit quantized self-attention map. We then develop an information rectification module (IRM) and a distribution guided distillation (DGD) scheme for fully quantized vision transformers (Q-ViT) to effectively eliminate such distortion, leading to a fully quantized ViTs. We evaluate our methods on popular DeiT and Swin backbones. Extensive experimental results show that our method achieves a much better performance than the prior arts. For example, our Q-ViT can theoretically accelerates the ViT-S by 6.14x and achieves about 80.9% Top-1 accuracy, even surpassing the full-precision counterpart by 1.0% on ImageNet dataset.

## [Monte Carlo Tree Search based Variable Selection for High Dimensional Bayesian Optimization](#)

- Lei Song · Ke Xue · Xiaobin Huang · Chao Qian
- abstract@[open-review](#): Bayesian optimization (BO) is a class of popular methods for expensive black-box optimization, and has been widely applied to many scenarios. However, BO suffers from the curse of dimensionality, and scaling it to high-dimensional problems is still a challenge. In this paper, we propose a variable selection method MCTS-VS based on Monte Carlo tree search (MCTS), to iteratively select and optimize a subset of variables. That is, MCTS-VS constructs a low-dimensional subspace via MCTS and optimizes in the subspace with any BO algorithm. We give a theoretical analysis of the general variable selection method to reveal how it can work. Experiments on high-dimensional synthetic functions and real-world problems (e.g., MuJoCo locomotion tasks) show that MCTS-VS equipped with a proper BO optimizer can achieve state-of-the-art performance.

## [Understanding the Failure of Batch Normalization for Transformers in NLP](#)

- Jiaxi Wang · Ji Wu · Lei Huang
- abstract@[open-review](#): Batch Normalization (BN) is a core and prevalent technique in accelerating the training of deep neural networks and improving the generalization on Computer Vision (CV) tasks. However, it fails to defend its position in Natural Language Processing (NLP), which is dominated by Layer Normalization (LN). In this paper, we are trying to answer why BN usually performs worse than LN in NLP tasks with Transformer models. We find that the inconsistency between training and inference of BN is the leading cause that results in the failure of BN in NLP. We define Training Inference Discrepancy (TID) to quantitatively measure this inconsistency and reveal that TID can indicate BN's performance, supported by extensive experiments, including image classification, neural machine translation, language modeling, sequence labeling, and text classification tasks. We find that BN can obtain much better test performance than LN when TID keeps small through training. To suppress the explosion of TID, we propose Regularized BN (RBN) that adds a simple regularization term to narrow the gap between batch statistics and population statistics of BN. RBN improves the performance of BN consistently and outperforms or is on par with LN on 17 out of 20 settings, including ten datasets and two common variants of Transformer.

## [Faster and Scalable Algorithms for Densest Subgraph and Decomposition](#)

- Elfarouk Harb · Kent Quanrud · Chandra Chekuri
- abstract@[open-review](#): We study the densest subgraph problem (DSG) and the densest subgraph local decomposition problem (DSG-LD) in undirected graphs. We also consider supermodular generalizations of these problems. For large scale graphs simple iterative algorithms perform much better in practice than theoretically fast algorithms based on network-flow or LP solvers. Boob et al [1] recently gave a fast iterative algorithm called Greedy++ for DSG. It was shown in [2] that it converges to a  $\$(1-\epsilon)\$$  relative approximation to the optimum density in  $\$O(\frac{1}{\epsilon^2} \frac{1}{\Delta(G)} \frac{\lambda^2}{\Delta(G)})\$$  iterations where  $\Delta(G)$  is the maximum degree and  $\lambda$  is the optimum density. Danisch et al. [3] gave an iterative algorithm based on the Frank-Wolfe algorithm for DSG-LD that takes  $\$O(\frac{m\Delta(G)}{\epsilon^2})\$$  iterations to converge to an  $\epsilon$ -additive approximate local decomposition vector  $\hat{v}$ , where  $m$  is number of edges in the graph. In this paper we give a new iterative algorithm for both problems that takes at most  $\$O(\sqrt{m\Delta(G)} / \epsilon)\$$  iterations to converge to an  $\epsilon$ -additive approximate local decomposition vector; each iteration can be implemented in  $\$O(m)\$$  time. We describe a fractional peeling technique which has strong empirical performance as well as theoretical guarantees. The algorithm is scalable and simple, and can be applied to graphs with hundreds of millions of edges. We test our algorithm on real and synthetic data sets and show that it provides a significant benefit over previous algorithms. The algorithm and analysis extends to hypergraphs.

## [Approximation with CNNs in Sobolev Space: with Applications to Classification](#)

- Jian Huang · GUOHUAO SHEN · Yuling Jiao · Yuanyuan Lin
- abstract@[open-review](#): We derive a novel approximation error bound with explicit prefactor for Sobolev-regular functions using deep convolutional neural networks (CNNs). The bound is non-asymptotic in terms of the network depth and filter lengths, in a rather flexible way. For Sobolev-regular functions which can be embedded into the  $H^s$  space, the prefactor of our error bound depends on the ambient dimension polynomially instead of exponentially as in most existing results, which is of independent interest. We also establish a new approximation result when the target function is supported on an approximate lower-dimensional manifold. We apply our results to establish non-asymptotic excess risk bounds for classification using CNNs with convex surrogate losses, including the cross-entropy loss, the hinge loss (SVM), the logistic loss, the exponential loss and the least squares loss. We show that the classification methods with CNNs can circumvent the curse of dimensionality if input data is supported on a neighborhood of a low-dimensional manifold.

## [On Robust Multiclass Learnability](#)

- Jingyuan Xu · Weiwei Liu
- abstract@[open-review](#): This work analyzes the robust learning problem in the multiclass setting. Under the framework of Probably Approximately Correct (PAC) learning, we first show that the graph dimension and the Natarajan dimension, which characterize the standard multiclass learnability, are no longer applicable in robust learning problem. We then generalize these notions to the robust learning setting, denoted as the adversarial graph dimension (AG-dimension) and the adversarial Natarajan dimension (AN-dimension). Upper and lower bounds of the sample complexity of robust multiclass learning are rigorously derived based on the AG-dimension and AN-dimension, respectively. Moreover, we calculate the AG-dimension and AN-dimension of the class of linear multiclass predictors, and show that the graph (Natarajan) dimension is of the same order as the AG(AN)-dimension. Finally, we prove that the AG-dimension and AN-dimension are not equivalent.

## [Robust Graph Structure Learning over Images via Multiple Statistical Tests](#)

- Yaohua Wang · Fangyi Zhang · Ming Lin · Senzhang Wang · Xiuyu Sun · Rong Jin
- abstract@[open-review](#): Graph structure learning aims to learn connectivity in a graph from data. It is particularly important for many computer vision related tasks since no explicit graph structure is available for images for most cases. A natural way to construct a graph among images is to treat each image as a node and assign pairwise image similarities as weights to corresponding edges. It is well known that pairwise similarities between images are sensitive to the noise in feature representations, leading to unreliable graph structures. We address this problem from the viewpoint of statistical tests. By viewing the feature vector of each node as an independent sample, the decision of whether creating an edge between two nodes based on their similarity in feature representation can be thought as a  $\$t$  statistical test. To improve the robustness in the decision of creating an edge, multiple samples are drawn and integrated by  $\$t$  statistical tests to generate a more reliable similarity measure, consequentially more reliable graph structure. The corresponding elegant matrix form named  $\$mathcal{B}\$$  is designed for efficiency. The effectiveness of multiple tests for graph structure learning is verified both theoretically and empirically on multiple clustering and ReID benchmark datasets.

## [Constrained Update Projection Approach to Safe Policy Optimization](#)

- Long Yang · Jiaming Ji · Juntao Dai · Linrui Zhang · Binbin Zhou · Pengfei Li · Yaodong Yang · Gang Pan
- abstract@[open-review](#): Safe reinforcement learning (RL) studies problems where an intelligent agent has to not only maximize reward but also avoid exploring unsafe areas. In this study, we propose CUP, a novel policy optimization method based on Constrained Update Projection framework that enjoys rigorous safety guarantee. Central to our CUP development is the newly proposed surrogate functions along with the performance bound. Compared to previous safe RL methods, CUP enjoys the benefits of 1) CUP generalizes the surrogate functions to generalized advantage estimator (GAE), leading to

strong empirical performance. 2) CUP unifies performance bounds, providing a better understanding for some existing algorithms; 3) CUP provides a non-convex implementation via only first-order optimizers, which does not require any strong approximation on the convexity of the objectives. To validate our CUP method, we compared CUP against a comprehensive list of safe RL baselines on a wide range of tasks. Experiments show the effectiveness of CUP both in terms of reward and safety constraint satisfaction.

## [Graph Coloring via Neural Networks for Haplotype Assembly and Viral Quasispecies Reconstruction](#)

- Hansheng Xue · Vaibhav Rajan · Yu Lin
- abstract@[open-review](#): Understanding genetic variation, e.g., through mutations, in organisms is crucial to unravel their effects on the environment and human health. A fundamental characterization can be obtained by solving the haplotype assembly problem, which yields the variation across multiple copies of chromosomes. Variations among fast evolving viruses that lead to different strains (called quasispecies) are also deciphered with similar approaches. In both these cases, high-throughput sequencing technologies that provide oversampled mixtures of large noisy fragments (reads) of genomes, are used to infer constituent components (haplotypes or quasispecies). The problem is harder for polyploid species where there are more than two copies of chromosomes. State-of-the-art neural approaches to solve this NP-hard problem do not adequately model relations among the reads that are important for deconvolving the input signal. We address this problem by developing a new method, called NeurHap, that combines graph representation learning with combinatorial optimization. Our experiments demonstrate the substantially better performance of NeurHap in real and synthetic datasets compared to competing approaches.

## [Momentum Adversarial Distillation: Handling Large Distribution Shifts in Data-Free Knowledge Distillation](#)

- Kien Do · Thai Hung Le · Dung Nguyen · Dang Nguyen · HARIPRIYA HARIKUMAR · Truyen Tran · Santu Rana · Svetha Venkatesh
- abstract@[open-review](#): Data-free Knowledge Distillation (DFKD) has attracted attention recently thanks to its appealing capability of transferring knowledge from a teacher network to a student network without using training data. The main idea is to use a generator to synthesize data for training the student. As the generator gets updated, the distribution of synthetic data will change. Such distribution shift could be large if the generator and the student are trained adversarially, causing the student to forget the knowledge it acquired at the previous steps. To alleviate this problem, we propose a simple yet effective method called Momentum Adversarial Distillation (MAD) which maintains an exponential moving average (EMA) copy of the generator and uses synthetic samples from both the generator and the EMA generator to train the student. Since the EMA generator can be considered as an ensemble of the generator's old versions and often undergoes a smaller change in updates compared to the generator, training on its synthetic samples can help the student recall the past knowledge and prevent the student from adapting too quickly to the new updates of the generator. Our experiments on six benchmark datasets including big datasets like ImageNet and Places365 demonstrate the superior performance of MAD over competing methods for handling the large distribution shift problem. Our method also compares favorably to existing DFKD methods and even achieves state-of-the-art results in some cases.

## [Assistive Teaching of Motor Control Tasks to Humans](#)

- Megha Srivastava · Erdem Biyik · Suvir Mirchandani · Noah Goodman · Dorsa Sadigh
- abstract@[open-review](#): Recent works on shared autonomy and assistive-AI technologies, such as assistive robotic teleoperation, seek to model and help human users with limited ability in a fixed task. However, these approaches often fail to account for humans' ability to adapt and eventually learn how to execute a control task themselves. Furthermore, in applications where it may be desirable for a human to intervene, these methods may have inhibited their ability to learn how to succeed with full self-control. In this paper, we focus on the problem of assistive teaching of motor control tasks such as parking a car or landing an aircraft. Despite their ubiquitous role in humans' daily activities and occupations, motor tasks are rarely taught in a uniform way due to their high complexity and variance. We propose an AI-assisted teaching algorithm that leverages skill discovery methods from reinforcement learning (RL) literature to (i) break down any motor control task into teachable skills, (ii) construct novel drill sequences, and (iii) individualize curricula to students with different capabilities. Through an extensive mix of synthetic and user studies on two motor control tasks - parking a car with a joystick and writing characters from the Balinese alphabet - we show that assisted teaching with skills improve student performance by around 40% compared to practicing full trajectories without skills, and practicing with individualized drills can result in up to 25% further improvement.

## [Uni\[MASK\]: Unified inference in sequential decision problems](#)

- Micah Carroll · Orr Paradise · Jessy Lin · Raluca Georgescu · Mingfei Sun · David Bignell · Stephanie Milani · Katja Hofmann · Matthew Hausknecht · Anca Dragan · Sam Devlin
- abstract@[open-review](#): Randomly masking and predicting word tokens has been a successful approach in pre-training language models for a variety of downstream tasks. In this work, we observe that the same idea also applies naturally to sequential decision making, where many well-studied tasks like behavior cloning, offline RL, inverse dynamics, and waypoint conditioning correspond to different sequence maskings over a sequence of states, actions, and returns. We introduce the UniMASK framework, which provides a unified way to specify models which can be trained on many different sequential decision making tasks. We show that a single UniMASK model is often capable of carrying out many tasks with performance similar to or better than single-task models. Additionally, after fine-tuning, our UniMASK models consistently outperform comparable single-task models.

## [Dynamic pricing and assortment under a contextual MNL demand](#)

- Noemie Perivier · Vineet Goyal
- abstract@[open-review](#): We consider dynamic multi-product pricing and assortment problems under an unknown demand over  $T$  periods, where in each period, the seller decides on the price for each product or the assortment of products to offer to a customer who chooses according to an unknown Multinomial Logit Model (MNL). Such problems arise in many applications, including online retail and advertising. We propose a randomized dynamic pricing policy based on a variant of the Online Newton Step algorithm (ONS) that achieves a  $\$O(d\sqrt{T}\log(T))\$$  regret guarantee under an adversarial arrival model. We also present a new optimistic algorithm for the adversarial MNL contextual bandits problem, which achieves a better dependency than the state-of-the-art algorithms in a problem-dependent constant  $\$\\kappa\$$  (potentially exponentially small). Our regret upper bound scales as  $\$\\tilde{O}(d\sqrt{\\kappa T} + \\log(T)/\\kappa)\$$ , which gives a stronger bound than the existing  $\$\\tilde{O}(d\sqrt{T}/\\kappa)\$$  guarantees.

## [Debiased Machine Learning without Sample-Splitting for Stable Estimators](#)

- Qizhao Chen · Vasilis Syrgkanis · Morgane Austern
- abstract@[open-review](#): Estimation and inference on causal parameters is typically reduced to a generalized method of moments problem, which involves auxiliary functions that correspond to solutions to a regression or classification problem. Recent line of work on debiased machine learning shows how one can use generic machine learning estimators for these auxiliary problems, while maintaining asymptotic normality and root-\$n\$ consistency of the target parameter of interest, while only requiring mean-squared-error guarantees from the auxiliary estimation algorithms. The literature typically requires that these auxiliary problems are fitted on a separate sample or in a cross-fitting manner. We show that when these auxiliary estimation algorithms satisfy natural leave-one-out stability properties, then sample splitting is not required. This allows for sample re-use, which can be beneficial in moderately sized sample regimes. For instance, we show that the stability properties that we propose are satisfied for ensemble bagged estimators, built via sub-sampling without replacement, a popular technique in machine learning practice.

## The Missing Invariance Principle found -- the Reciprocal Twin of Invariant Risk Minimization

- Dongsung Huh · Avinash Baidya
- abstract@[open-review](#): Machine learning models often generalize poorly to out-of-distribution (OOD) data as a result of relying on features that are spuriously correlated with the label during training. Recently, the technique of Invariant Risk Minimization (IRM) was proposed to learn predictors that only use invariant features by conserving the feature-conditioned label expectation  $\mathbb{E}_e[y|f(x)]$  across environments. However, more recent studies have demonstrated that IRM-v1, a practical version of IRM, can fail in various task settings. Here, we identify a fundamental flaw of IRM formulation that causes the failure. We then introduce a complementary notion of invariance, MRI, based on conserving the label-conditioned feature expectation  $\mathbb{E}_e[f(x)|y]$  across environments, which is free of this flaw. Further, we introduce a simplified, practical version of the MRI formulation called MRI-v1. We note that this constraint is convex which confers it with an advantage over IRM-v1, which imposes non-convex constraints. We prove that in a general linear problem setting, MRI-v1 can guarantee invariant predictors given sufficient environments. We also empirically demonstrate that MRI strongly out-performs IRM and consistently achieves near-optimal OOD generalization in image-based nonlinear problems.

## Capturing Graphs with Hypo-Elliptic Diffusions

- Csaba Toth · Darrick Lee · Celia Hacker · Harald Oberhauser
- abstract@[open-review](#): Convolutional layers within graph neural networks operate by aggregating information about local neighbourhood structures; one common way to encode such substructures is through random walks. The distribution of these random walks evolves according to a diffusion equation defined using the graph Laplacian. We extend this approach by leveraging classic mathematical results about hypo-elliptic diffusions. This results in a novel tensor-valued graph operator, which we call the hypo-elliptic graph Laplacian. We provide theoretical guarantees and efficient low-rank approximation algorithms. In particular, this gives a structured approach to capture long-range dependencies on graphs that is robust to pooling. Besides the attractive theoretical properties, our experiments show that this method competes with graph transformers on datasets requiring long-range reasoning but scales only linearly in the number of edges as opposed to quadratically in nodes.

## Private and Communication-Efficient Algorithms for Entropy Estimation

- Gecia Bravo-Hermsdorff · RÃ©bert Busa-Fekete · Mohammad Ghavamzadeh · Andres Munoz Medina · Umar Syed
- abstract@[open-review](#): Modern statistical estimation is often performed in a distributed setting where each sample belongs to single user who shares their data with a central server. Users are typically concerned with preserving the privacy of their sample, and also with minimizing the amount of data they must transmit to the server. We give improved private and communication-efficient algorithms for estimating several popular measures of the entropy of a distribution. All of our algorithms have constant communication cost and satisfy local differential privacy. For a joint distribution on several variables whose conditional independence graph is a tree, we describe algorithms for estimating Shannon entropy that require a number of samples that is linear in the number of variables, compared to the quadratic sample complexity of prior work. We also describe an algorithm for estimating Gini entropy whose sample complexity has no dependence on the support size of the distribution and can be implemented using a single round of concurrent communication between the users and the server, while the previously best-known algorithm has high communication cost and requires the server to facilitate interaction between the users. Finally, we describe an algorithm for estimating collision entropy that matches the space and sample complexity of the best known algorithm but generalizes it to the private and communication-efficient setting.

## Sequential Hypothesis Tests of Multinomial Count Data

- Michael Lindon · Alan Malek
- abstract@[open-review](#): Sequential hypothesis tests of equality and contrasts among arbitrarily many time-inhomogeneous Bernoulli and Poisson counting processes are constructed based on a new sequential test of multinomial point hypotheses. For multinomial, Bernoulli and Poisson counting processes we provide confidence sequences for the probability vector, all contrasts in log-probabilities and all contrasts in log-intensities, respectively. Combined with sequential p-values, these provide an "always-valid" framework for inference, controlling the Type I probability under optional stopping/continuation and continuous monitoring. These methods are demonstrated with three applications relevant to online controlled experiments: sample ratio mismatch testing, conversion rate optimization, and software canary testing.

## The least-control principle for learning at equilibrium

- Alexander Meulemans · Nicolas Zucchet · Seijin Kobayashi · Johannes von Oswald · JoÃ£o Sacramento
- abstract@[open-review](#): Equilibrium systems are a powerful way to express neural computations. As special cases, they include models of great current interest in both neuroscience and machine learning, such as equilibrium recurrent neural networks, deep equilibrium models, or meta-learning. Here, we present a new principle for learning such systems with a temporally- and spatially-local rule. Our principle casts learning as a least-control problem, where we first introduce an optimal controller to lead the system towards a solution state, and then define learning as reducing the amount of control needed to reach such a state. We show that incorporating learning signals within a dynamics as an optimal control enables transmitting credit assignment information in previously unknown ways, avoids storing intermediate states in memory, and does not rely on infinitesimal learning signals. In practice, our principle leads to strong performance matching that of leading gradient-based learning methods when applied to an array of problems involving recurrent neural networks and meta-learning. Our results shed light on how the brain might learn and offer new ways of approaching a broad class of machine learning problems.

## Reinforcement Learning in a Birth and Death Process: Breaking the Dependence on the State Space

- Jonatha Anselmi · Bruno Gaujal · Louis-SÃ©bastien Rebuffi
- abstract@[open-review](#): In this paper, we revisit the regret of undiscounted reinforcement learning in MDPs with a birth and death structure. Specifically, we consider a controlled queue with impatient jobs and the main objective is to optimize a trade-off between energy consumption and user-perceived performance. Within this setting, the diameter  $D$  of the MDP is  $\Omega(S^S)$ , where  $S$  is the number of states. Therefore, the existing lower and upper bounds on the regret at time  $T$ , of order  $O(\sqrt{DSAT})$  for MDPs with  $S$  states and  $A$  actions, may suggest that reinforcement learning is inefficient here. In our main result however, we exploit the structure of our MDPs to show that the regret of a slightly-tweaked version of the classical learning algorithm UCRL2 is in fact upper bounded by  $\tilde{O}(\sqrt{E_2 AT})$  where  $E_2$  is a weighted second moment of the stationary measure of a reference policy. Importantly,  $E_2$  is bounded independently of  $S$ . Thus, our bound is asymptotically independent of the number of states and of the diameter. This result is based on a careful study of the number of visits performed by the learning algorithm to the states of the MDP, which is highly non-uniform.

## Ordered Subgraph Aggregation Networks

- Chendi Qian · Gaurav Rattan · Floris Geerts · Mathias Niepert · Christopher Morris
- abstract@[open-review](#): Recently, many subgraph-enhanced graph neural networks emerged, provably boosting the expressive power of standard (message-passing) graph neural networks. However, there is a limited understanding of how these approaches relate to each other and the Weisfeiler-Leman hierarchy. Further, current approaches either use all subgraphs of a given size, sample them uniformly at random, or use hand-crafted heuristics to

select them, oblivious to the given data distribution. Here, we offer a unified way to study such architectures by introducing a theoretical framework and extending the known expressivity results of subgraph-enhanced graph neural networks. That is, we show that increasing subgraph size always increases expressive power and develop a better understanding of their limitations by relating them to the established  $\text{WL}$  hierarchy. In addition, we explore different approaches for sampling subgraphs using state-of-the-art data-driven methods for backpropagating through discrete structures using perturbation-based implicit differentiation. Empirically, we study the predictive performance of different subgraph-enhanced graph neural networks, showing that our data-driven architectures increase prediction accuracy on standard benchmark datasets compared to non-data-driven subgraph-enhanced graph neural networks while vastly reducing computation time.

## [Adapting to Domain Shift by Meta-Distillation from Mixture-of-Experts](#)

- Tao Zhong · Zhixiang Chi · Li Gu · Yang Wang · Yuanhao Yu · Jin Tang
- abstract@[open-review](#): In this paper, we tackle the problem of domain shift. Most existing methods perform training on multiple source domains using a single model, and the same trained model is used on all unseen target domains. Such solutions are sub-optimal as each target domain exhibits its own speciality, which is not adapted. Furthermore, expecting the single-model training to learn extensive knowledge from the multiple source domains is counterintuitive. The model is more biased to learning only domain-invariant features and may result in negative knowledge transfer. In this work, we propose a novel framework for unsupervised test-time adaptation, which is formulated as a knowledge distillation process to address domain shift. Specifically, we incorporate with Mixture-of-Experts (MoE) as teachers, where each expert is separately trained on different source domains to maximize their speciality. Given a test-time target domain, a small set of unlabeled data is sampled to query the knowledge from MoE. As the source domains are correlated to the target domains, a transformer-based aggregator then combines the domain knowledge by examining the interconnection among them. The output is treated as a supervision signal to adapt a student prediction network toward the target domain. We further employ meta-learning to enforce the aggregator to distill positive knowledge and the student network to achieve fast adaptation. Extensive experiments demonstrate that the proposed method outperforms the state-of-the-art and validates the effectiveness of each proposed component.

## [Invariant and Transportable Representations for Anti-Causal Domain Shifts](#)

- Yibo Jiang · Victor Veitch
- abstract@[open-review](#): Real-world classification problems must contend with domain shift, the (potential) mismatch between the domain where a model is deployed and the domain(s) where the training data was gathered. Methods to handle such problems must specify what structure is held in common between the domains and what is allowed to vary. A natural assumption is that causal (structural) relationships are invariant in all domains. Then, it is tempting to learn a predictor for label  $Y$  that depends only on its causal parents. However, many real-world problems are "anti-causal" in the sense that  $Y$  is a cause of the covariates  $X$ —in this case,  $Y$  has no causal parents and the naive causal invariance is useless. In this paper, we study representation learning under a particular notion of domain shift that both respects causal invariance and that naturally handles the "anti-causal" structure. We show how to leverage the shared causal structure of the domains to learn a representation that both admits an invariant predictor and that also allows fast adaptation in new domains. The key is to translate causal assumptions into learning principles that disentangle "invariant" and "non-stable" features. Experiments on both synthetic and real-world data demonstrate the effectiveness of the proposed learning algorithm.

## [Model Preserving Compression for Neural Networks](#)

- Jerry Chee · Megan Renz · Anil Damle · Christopher De Sa
- abstract@[open-review](#): After training complex deep learning models, a common task is to compress the model to reduce compute and storage demands. When compressing, it is desirable to preserve the original model's per-example decisions (e.g., to go beyond top-1 accuracy or preserve robustness), maintain the network's structure, automatically determine per-layer compression levels, and eliminate the need for fine tuning. No existing compression methods simultaneously satisfy these criteria—we introduce a principled approach that does by leveraging interpolative decompositions. Our approach simultaneously selects and eliminates channels (analogously, neurons), then constructs an interpolation matrix that propagates a correction into the next layer, preserving the network's structure. Consequently, our method achieves good performance even without fine tuning and admits theoretical analysis. Our theoretical generalization bound for a one layer network lends itself naturally to a heuristic that allows our method to automatically choose per-layer sizes for deep networks. We demonstrate the efficacy of our approach with strong empirical performance on a variety of tasks, models, and datasets—from simple one-hidden-layer networks to deep networks on ImageNet.

## [Non-identifiability and the Blessings of Misspecification in Models of Molecular Fitness](#)

- Eli Weinstein · Alan Amin · Jonathan Frazer · Debora Marks
- abstract@[open-review](#): Understanding the consequences of mutation for molecular fitness and function is a fundamental problem in biology. Recently, generative probabilistic models have emerged as a powerful tool for estimating fitness from evolutionary sequence data, with accuracy sufficient to predict both laboratory measurements of function and disease risk in humans, and to design novel functional proteins. Existing techniques rest on an assumed relationship between density estimation and fitness estimation, a relationship that we interrogate in this article. We prove that fitness is not identifiable from observational sequence data alone, placing fundamental limits on our ability to disentangle fitness landscapes from phylogenetic history. We show on real datasets that perfect density estimation in the limit of infinite data would, with high confidence, result in poor fitness estimation; current models perform accurate fitness estimation because of, not despite, misspecification. Our results challenge the conventional wisdom that bigger models trained on bigger datasets will inevitably lead to better fitness estimation, and suggest novel estimation strategies going forward.

## [Meta-learning for Feature Selection with Hilbert-Schmidt Independence Criterion](#)

- Atsutoshi Kumagai · Tomoharu Iwata · Yasutoshi Ida · Yasuhiro Fujiwara
- abstract@[open-review](#): We propose a meta-learning method for feature selection that can select relevant features given a small number of labeled instances. Existing methods require many labeled instances for accurate feature selection. However, sufficient instances are often unavailable. We use labeled instances in multiple related tasks to alleviate the lack of labeled instances in a target task. To measure the dependency between each feature and label, we use the Hilbert-Schmidt Independence Criterion, which is a kernel-based independence measure. By modeling the kernel functions with neural networks that take a few labeled instances in a task as input, we can encode the task-specific information to the kernels such that the kernels are appropriate for the task. Feature selection with such kernels is performed by using iterative optimization methods, in which each update step is obtained as a closed-form. This formulation enables us to directly and efficiently minimize the expected test error on features selected by a small number of labeled instances. We experimentally demonstrate that the proposed method outperforms existing feature selection methods.

## [Generalization Gap in Amortized Inference](#)

- Mingtian Zhang · Peter Hayes · David Barber
- abstract@[open-review](#): The ability of likelihood-based probabilistic models to generalize to unseen data is central to many machine learning applications such as lossless compression. In this work, we study the generalizations of a popular class of probabilistic models - the Variational Auto-Encoder (VAE). We point out the two generalization gaps that can affect the generalization ability of VAEs and show that the over-fitting phenomenon is usually dominated by the amortized inference network. Based on this observation we propose a new training objective, inspired by the classic wake-sleep

algorithm, to improve the generalizations properties of amortized inference. We also demonstrate how it can improve generalization performance in the context of image modeling and lossless compression.

## [On the Safety of Interpretable Machine Learning: A Maximum Deviation Approach](#)

- Dennis Wei · Rahul Nair · Amit Dhurandhar · Kush Varshney · Elizabeth Daly · Moninder Singh
- abstract@[open-review](#): Interpretable and explainable machine learning has seen a recent surge of interest. We focus on safety as a key motivation behind the surge and make the relationship between safety and interpretability more quantitative. Toward assessing safety, we introduce the concept of maximum deviation via an optimization problem to find the largest deviation of a supervised learning model from a reference model regarded as safe. We then show how interpretability facilitates this safety assessment. For models including decision trees, generalized linear and additive models, the maximum deviation can be computed exactly and efficiently. For tree ensembles, which are not regarded as interpretable, discrete optimization techniques can still provide informative bounds. For a broader class of piecewise Lipschitz functions, we leverage the multi-armed bandit literature to show that interpretability produces tighter (regret) bounds on the maximum deviation. We present case studies, including one on mortgage approval, to illustrate our methods and the insights about models that may be obtained from deviation maximization.

## [Graph Agnostic Estimators with Staggered Rollout Designs under Network Interference](#)

- Mayleen Cortez · Matthew Eichhorn · Christina Yu
- abstract@[open-review](#): Randomized experiments are widely used to estimate causal effects across a variety of domains. However, classical causal inference approaches rely on critical independence assumptions that are violated by network interference, when the treatment of one individual influences the outcomes of others. All existing approaches require at least approximate knowledge of the network, which may be unavailable and costly to collect. We consider the task of estimating the total treatment effect (TTE), or the average difference between the outcomes when the whole population is treated versus when the whole population is untreated. By leveraging a staggered rollout design, in which treatment is incrementally given to random subsets of individuals, we derive unbiased estimators for TTE that do not rely on any prior structural knowledge of the network, as long as the network interference effects are constrained to low-degree interactions among neighbors of an individual. We derive bounds on the variance of the estimators, and we show in experiments that our estimator performs well against baselines on simulated data. Central to our theoretical contribution is a connection between staggered rollout observations and polynomial extrapolation.

## [Kernel Interpolation with Sparse Grids](#)

- Mohit Yadav · Daniel Sheldon · Cameron Musco
- abstract@[open-review](#): Structured kernel interpolation (SKI) accelerates Gaussian processes (GP) inference by interpolating the kernel covariance function using a dense grid of inducing points, whose corresponding kernel matrix is highly structured and thus amenable to fast linear algebra. Unfortunately, SKI scales poorly in the dimension of the input points, since the dense grid size grows exponentially with the dimension. To mitigate this issue, we propose the use of sparse grids within the SKI framework. These grids enable accurate interpolation, but with a number of points growing more slowly with dimension. We contribute a novel nearly linear time matrix-vector multiplication algorithm for the sparse grid kernel matrix. We also describe how sparse grids can be combined with an efficient interpolation scheme based on simplicial complexes. With these modifications, we demonstrate that SKI can be scaled to higher dimensions while maintaining accuracy, for both synthetic and real datasets.

## [Value Function Decomposition for Iterative Design of Reinforcement Learning Agents](#)

- James MacGlashan · Evan W Archer · Alisa Devlic · Takuma Seno · Craig Sherstan · Peter Wurman · Peter Stone
- abstract@[open-review](#): Designing reinforcement learning (RL) agents is typically a difficult process that requires numerous design iterations. Learning can fail for a multitude of reasons and standard RL methods provide too few tools to provide insight into the exact cause. In this paper, we show how to integrate value decomposition into a broad class of actor-critic algorithms and use it to assist in the iterative agent-design process. Value decomposition separates a reward function into distinct components and learns value estimates for each. These value estimates provide insight into an agent's learning and decision-making process and enable new training methods to mitigate common problems. As a demonstration, we introduce SAC-D, a variant of soft actor-critic (SAC) adapted for value decomposition. SAC-D maintains similar performance to SAC, while learning a larger set of value predictions. We also introduce decomposition-based tools that exploit this information, including a new reward influence metric, which measures each reward component's effect on agent decision-making. Using these tools, we provide several demonstrations of decomposition's use in identifying and addressing problems in the design of both environments and agents. Value decomposition is broadly applicable and easy to incorporate into existing algorithms and workflows, making it a powerful tool in an RL practitioner's toolbox.

## [Gradient Estimation with Discrete Stein Operators](#)

- Jiaxin Shi · Yuhao Zhou · Jessica Hwang · Michalis Titsias · Lester Mackey
- abstract@[open-review](#): Gradient estimation---approximating the gradient of an expectation with respect to the parameters of a distribution---is central to the solution of many machine learning problems. However, when the distribution is discrete, most common gradient estimators suffer from excessive variance. To improve the quality of gradient estimation, we introduce a variance reduction technique based on Stein operators for discrete distributions. We then use this technique to build flexible control variates for the REINFORCE leave-one-out estimator. Our control variates can be adapted online to minimize variance and do not require extra evaluations of the target function. In benchmark generative modeling tasks such as training binary variational autoencoders, our gradient estimator achieves substantially lower variance than state-of-the-art estimators with the same number of function evaluations.

## [Composite Feature Selection Using Deep Ensembles](#)

- Alexander Norcliffe · Fergus Imrie · Pietro Lī · Mihaela van der Schaar
- abstract@[open-review](#): In many real world problems, features do not act alone but in combination with each other. For example, in genomics, diseases might not be caused by any single mutation but require the presence of multiple mutations. Prior work on feature selection either seeks to identify individual features or can only determine relevant groups from a predefined set. We investigate the problem of discovering groups of predictive features without predefined grouping. To do so, we define predictive groups in terms of linear and non-linear interactions between features. We introduce a novel deep learning architecture that uses an ensemble of feature selection models to find predictive groups, without requiring candidate groups to be provided. The selected groups are sparse and exhibit minimum overlap. Furthermore, we propose a new metric to measure similarity between discovered groups and the ground truth. We test our model on multiple synthetic tasks, semi-synthetic chemistry datasets and image datasets to demonstrate its utility.

## [Explain My Surprise: Learning Efficient Long-Term Memory by predicting uncertain outcomes](#)

- Artyom Sorokin · Nazar Buzun · Leonid Pugachev · Mikhail Burtsev
- abstract@[open-review](#): In many sequential tasks, a model needs to remember relevant events from the distant past to make correct predictions. Unfortunately, a straightforward application of gradient based training requires intermediate computations to be stored for every element of a sequence. This requires prohibitively large compute memory if a sequence consists of thousands or even millions elements, and as a result, makes learning of very long-term dependencies infeasible. However, the majority of sequence elements can usually be predicted by taking into account only temporally local

information. On the other hand, predictions affected by long-term dependencies are sparse and characterized by high uncertainty given only local information. We propose MemUP, a new training method that allows to learn long-term dependencies without backpropagating gradients through the whole sequence at a time. This method can be potentially applied to any gradient based sequence learning. MemUP implementation for recurrent architectures shows performances better or comparable to baselines while requiring significantly less compute memory.

## [Uplifting Bandits](#)

- Yu-Guan Hsieh · Shiva Kasiviswanathan · Branislav Kveton
- abstract@[open-review](#): We introduce a new multi-armed bandit model where the reward is a sum of multiple random variables, and each action only alters the distributions of some of these variables. Upon taking an action, the agent observes the realizations of all variables. This model is motivated by marketing campaigns and recommender systems, where the variables represent outcomes on individual customers, such as clicks. We propose UCB-style algorithms that estimate the uplifts of the actions over a baseline. We study multiple variants of the problem, including when the baseline and affected variables are unknown, and prove sublinear regret bounds for all of these. In addition, we provide regret lower bounds that justify the necessity of our modeling assumptions. Experiments on synthetic and real-world datasets demonstrate the benefit of methods that estimate the uplifts over policies that do not use this structure.

## [Off-Policy Evaluation with Policy-Dependent Optimization Response](#)

- Wenshuo Guo · Michael Jordan · Angela Zhou
- abstract@[open-review](#): The intersection of causal inference and machine learning for decision-making is rapidly expanding, but the default decision criterion remains an average of individual causal outcomes across a population. In practice, various operational restrictions ensure that a decision-maker's utility is not realized as an average but rather as an output of a downstream decision-making problem (such as matching, assignment, network flow, minimizing predictive risk). In this work, we develop a new framework for off-policy evaluation with policy-dependent linear optimization responses: causal outcomes introduce stochasticity in objective function coefficients. Under this framework, a decision-maker's utility depends on the policy-dependent optimization, which introduces a fundamental challenge of optimization bias even for the case of policy evaluation. We construct unbiased estimators for the policy-dependent estimand by a perturbation method, and discuss asymptotic variance properties for a set of adjusted plug-in estimators. Lastly, attaining unbiased policy evaluation allows for policy optimization: we provide a general algorithm for optimizing causal interventions. We corroborate our theoretical results with numerical simulations.

## [Safe Opponent-Exploitation Subgame Refinement](#)

- Mingyang Liu · Chengjie Wu · Qihan Liu · Yansen Jing · Jun Yang · Pingzhong Tang · Chongjie Zhang
- abstract@[open-review](#): In zero-sum games, an NE strategy tends to be overly conservative confronted with opponents of limited rationality, because it does not actively exploit their weaknesses. From another perspective, best responding to an estimated opponent model is vulnerable to estimation errors and lacks safety guarantees. Inspired by the recent success of real-time search algorithms in developing superhuman AI, we investigate the dilemma of safety and opponent exploitation and present a novel real-time search framework, called Safe Exploitation Search (SES), which continuously interpolates between the two extremes of online strategy refinement. We provide SES with a theoretically upper-bounded exploitability and a lower-bounded evaluation performance. Additionally, SES enables computationally efficient online adaptation to a possibly updating opponent model, while previous safe exploitation methods have to recompute for the whole game. Empirical results show that SES significantly outperforms NE baselines and previous algorithms while keeping exploitability low at the same time.

## [Bezier Gaussian Processes for Tall and Wide Data](#)

- Martin Jørgensen · Michael A Osborne
- abstract@[open-review](#): Modern approximations to Gaussian processes are suitable for tall data", with a cost that scales well in the number of observations, but under-performs on wide data", scaling poorly in the number of input features. That is, as the number of input features grows, good predictive performance requires the number of summarising variables, and their associated cost, to grow rapidly. We introduce a kernel that allows the number of summarising variables to grow exponentially with the number of input features, but requires only linear cost in both number of observations and input features. This scaling is achieved through our introduction of the ``Bezier buttress'', which allows approximate inference without computing matrix inverses or determinants. We show that our kernel has close similarities to some of the most used kernels in Gaussian process regression, and empirically demonstrate the kernel's ability to scale to both tall and wide datasets.

## [AgraSSt: Approximate Graph Stein Statistics for Interpretable Assessment of Implicit Graph Generators](#)

- Wenkai Xu · Gesine D Reinert
- abstract@[open-review](#): We propose and analyse a novel statistical procedure, coined AgraSSt, to assess the quality of graph generators which may not be available in explicit forms. In particular, AgraSSt can be used to determine whether a learned graph generating process is capable of generating graphs which resemble a given input graph. Inspired by Stein operators for random graphs, the key idea of AgraSSt is the construction of a kernel discrepancy based on an operator obtained from the graph generator. AgraSSt can provide interpretable criticisms for a graph generator training procedure and help identify reliable sample batches for downstream tasks. We give theoretical guarantees for a broad class of random graph models. We provide empirical results on both synthetic input graphs with known graph generation procedures, and real-world input graphs that the state-of-the-art (deep) generative models for graphs are trained on.

## [SAPD+: An Accelerated Stochastic Method for Nonconvex-Concave Minimax Problems](#)

- Xuan Zhang · Necdet Serhat Aybat · Mert Gurbuzbalaban
- abstract@[open-review](#): We propose a new stochastic method SAPD+ for solving nonconvex-concave minimax problems of the form  $\min \max \mathcal{L}(x, y) = f(x) + \Phi(x, y) - g(y)$ , where  $f, g$  are closed convex and  $\Phi(x, y)$  is a smooth function that is weakly convex in  $x$ , (strongly) concave in  $y$ . For both strongly concave and merely concave settings, SAPD+ achieves the best known oracle complexities of  $\mathcal{O}(L\kappa_y\epsilon^{-4})$  and  $\mathcal{O}(L^3\epsilon^{-6})$ , respectively, without assuming compactness of the problem domain, where  $\kappa_y$  is the condition number, and  $L$  is the Lipschitz constant. We also propose SAPD+ with variance reduction, which enjoys the best known oracle complexity of  $\mathcal{O}(L\kappa_y^2\epsilon^{-3})$  for weakly convex-strongly concave setting. We demonstrate the efficiency of SAPD+ on a distributionally robust learning problem with a weakly convex cost and also on a multi-class classification problem in deep learning.

## [Multi-layer State Evolution Under Random Convolutional Design](#)

- Max Daniels · Cedric Gerbelot · Florent Krzakala · Lenka Zdeborová
- abstract@[open-review](#): Signal recovery under generative neural network priors has emerged as a promising direction in statistical inference and computational imaging. Theoretical analysis of reconstruction algorithms under generative priors is, however, challenging. For generative priors with fully connected layers and Gaussian i.i.d. weights, this was achieved by the multi-layer approximate message (ML-AMP) algorithm via a rigorous state evolution. However, practical generative priors are typically convolutional, allowing for computational benefits and inductive biases, and so the Gaussian

i.i.d. weight assumption is very limiting. In this paper, we overcome this limitation and establish the state evolution of ML-AMP for random convolutional layers. We prove in particular that random convolutional layers belong to the same universality class as Gaussian matrices. Our proof technique is of an independent interest as it establishes a mapping between convolutional matrices and spatially coupled sensing matrices used in coding theory.

## [Biological Learning of Irreducible Representations of Commuting Transformations](#)

- Alexander Genkin · David Lipshutz · Siavash Golkar · Tiberiu Tesileanu · Dmitri Chklovskii
- abstract@[open-review](#): A challenge in neuroscience is to understand the neural mechanisms underlying the brain's remarkable ability to learn and detect transformations of objects. Cohen and Welling proposed a mathematical framework for learning efficient representations of transformations based on commutative Lie groups. In particular, they used a decomposition of a commutative Lie group into 2-dimensional irreducible representations. We explore the possibility that the brain uses a similar learning paradigm. Specifically, we propose bio-inspired mechanisms for learning these irreducible representations. We derive two algorithms for learning commutative groups from sequences of transformed images. The two approaches differ in their algorithmic foundations --- one is based on SVD of the anti-symmetrized outer product of consecutive frames and the other based on PCA of the difference between consecutive frames. The two algorithms lead to different neural network realizations --- the former uses a one layer network while the latter uses a two layer network. Both networks work in an online setting and are able to recover the results from Cohen and Welling, which focus on rotations, as well as learn more generic transformation groups. Our circuits suggest patterns that can be searched for in nascent connectomics and physiology datasets.

## [Group GAN](#)

- Ali Seyfi · Jean-Francois Rajotte · Raymond Ng
- abstract@[open-review](#): Generating multivariate time series is a promising approach for sharing sensitive data in many medical, financial, and IoT applications. A common type of multivariate time series originates from a single source such as the biometric measurements from a medical patient. This leads to complex dynamical patterns between individual time series that are hard to learn by typical generation models such as GANs. There is valuable information in those patterns that machine learning models can use to better classify, predict or perform other downstream tasks. We propose a novel framework that takes time series' common origin into account and favors inter-channel relationships preservation. The two key points of our method are: 1) the individual time series are generated from a common point in latent space and 2) a central discriminator favors the preservation of inter-channel dynamics. We demonstrate empirically that our method helps preserve channel correlations and that our synthetic data performs very well downstream tasks with medical and financial data.

## [SInGE: Sparsity via Integrated Gradients Estimation of Neuron Relevance](#)

- Edouard YVINEC · Arnaud Dapogny · Matthieu Cord · Kevin Bailly
- abstract@[open-review](#): The leap in performance in state-of-the-art computer vision methods is attributed to the development of deep neural networks. However it often comes at a computational price which may hinder their deployment. To alleviate this limitation, structured pruning is a well known technique which consists in removing channels, neurons or filters, and is commonly applied in order to produce more compact models. In most cases, the computations to remove are selected based on a relative importance criterion. At the same time, the need for explainable predictive models has risen tremendously and motivated the development of robust attribution methods that highlight the relative importance of pixels of an input image or feature map. In this work, we discuss the limitations of existing pruning heuristics, among which magnitude and gradient-based methods. We draw inspiration from attribution methods to design a novel integrated gradient pruning criterion, in which the relevance of each neuron is defined as the integral of the gradient variation on a path towards this neuron removal. Furthermore, We propose an entwined DNN pruning and fine-tuning flowchart to better preserve DNN accuracy while removing parameters. We show through extensive validation on several datasets, architectures as well as pruning scenarios that the proposed method, dubbed SInGE, significantly outperforms existing state-of-the-art DNN pruning methods.

## [A Closer Look at Prototype Classifier for Few-shot Image Classification](#)

- Mingcheng Hou · Issei Sato
- abstract@[open-review](#): The prototypical network is a prototype classifier based on meta-learning and is widely used for few-shot learning because it classifies unseen examples by constructing class-specific prototypes without adjusting hyper-parameters during meta-testing. Interestingly, recent research has attracted a lot of attention, showing that training a new linear classifier, which does not use a meta-learning algorithm, performs comparably with the prototypical network. However, the training of a new linear classifier requires the retraining of the classifier every time a new class appears. In this paper, we analyze how a prototype classifier works equally well without training a new linear classifier or meta-learning. We experimentally find that directly using the feature vectors, which is extracted by using standard pre-trained models to construct a prototype classifier in meta-testing, does not perform as well as the prototypical network and training new linear classifiers on the feature vectors of pre-trained models. Thus, we derive a novel generalization bound for a prototypical classifier and show that the transformation of a feature vector can improve the performance of prototype classifiers. We experimentally investigate several normalization methods for minimizing the derived bound and find that the same performance can be obtained by using the L2 normalization and minimizing the ratio of the within-class variance to the between-class variance without training a new classifier or meta-learning.

## [Deep invariant networks with differentiable augmentation layers](#)

- Cédric ROMMEL · Thomas Moreau · Alexandre Gramfort
- abstract@[open-review](#): Designing learning systems which are invariant to certain data transformations is critical in machine learning. Practitioners can typically enforce a desired invariance on the trained model through the choice of a network architecture, e.g. using convolutions for translations, or using data augmentation. Yet, enforcing true invariance in the network can be difficult, and data invariances are not always known a priori. State-of-the-art methods for learning data augmentation policies require held-out data and are based on bilevel optimization problems, which are complex to solve and often computationally demanding. In this work we investigate new ways of learning invariances only from the training data. Using learnable augmentation layers built directly in the network, we demonstrate that our method is very versatile. It can incorporate any type of differentiable augmentation and be applied to a broad class of learning problems beyond computer vision. We provide empirical evidence showing that our approach is easier and faster to train than modern automatic data augmentation techniques based on bilevel optimization, while achieving comparable results. Experiments show that while the invariances transferred to a model through automatic data augmentation are limited by the model expressivity, the invariance yielded by our approach is insensitive to it by design.

## [Periodic Graph Transformers for Crystal Material Property Prediction](#)

- Keqiang Yan · Yi Liu · Yuchao Lin · Shuiwang Ji
- abstract@[open-review](#): We consider representation learning on periodic graphs encoding crystal materials. Different from regular graphs, periodic graphs consist of a minimum unit cell repeating itself on a regular lattice in 3D space. How to effectively encode these periodic structures poses unique challenges not present in regular graph representation learning. In addition to being  $E(3)$  invariant, periodic graph representations need to be periodic invariant. That is, the learned representations should be invariant to shifts of cell boundaries as they are artificially imposed. Furthermore, the periodic repeating patterns need to be captured explicitly as lattices of different sizes and orientations may correspond to different materials. In this work, we propose a transformer architecture, known as Matformer, for periodic graph representation learning. Our Matformer is designed to be invariant to periodicity and can capture repeating patterns explicitly. In particular, Matformer encodes periodic patterns by efficient use of geometric distances

between the same atoms in neighboring cells. Experimental results on multiple common benchmark datasets show that our Matformer outperforms baseline methods consistently. In addition, our results demonstrate the importance of periodic invariance and explicit repeating pattern encoding for crystal representation learning.

## Unsupervised Cross-Task Generalization via Retrieval Augmentation

- Bill Yuchen Lin · Kangmin Tan · Chris Miller · Beiwen Tian · Xiang Ren
- abstract@[open-review](#): Humans can perform unseen tasks by recalling relevant skills that are acquired previously and then generalizing them to the target tasks, even if there is no supervision at all. In this paper, we aim to improve such cross-task generalization ability of massive multi-task language models such as T0 (Sanh et al., 2021) in an unsupervised setting. We propose a retrieval-augmentation method named ReCross that takes a few unlabelled examples as queries to retrieve a small subset of upstream data and uses them to update the multi-task model for better generalization. Our empirical results show that the proposed ReCross consistently outperforms non-retrieval baselines by a significant margin.

## Elucidating the Design Space of Diffusion-Based Generative Models

- Tero Karras · Miika Aittala · Timo Aila · Samuli Laine
- abstract@[open-review](#): We argue that the theory and practice of diffusion-based generative models are currently unnecessarily convoluted and seek to remedy the situation by presenting a design space that clearly separates the concrete design choices. This lets us identify several changes to both the sampling and training processes, as well as preconditioning of the score networks. Together, our improvements yield new state-of-the-art FID of 1.79 for CIFAR-10 in a class-conditional setting and 1.97 in an unconditional setting, with much faster sampling (35 network evaluations per image) than prior designs. To further demonstrate their modular nature, we show that our design changes dramatically improve both the efficiency and quality obtainable with pre-trained score networks from previous work, including improving the FID of an existing ImageNet-64 model from 2.07 to near-SOTA 1.55.

## Optimizing Relevance Maps of Vision Transformers Improves Robustness

- Hila Chefer · Idan Schwartz · Lior Wolf
- abstract@[open-review](#): It has been observed that visual classification models often rely mostly on the image background, neglecting the foreground, which hurts their robustness to distribution changes. To alleviate this shortcoming, we propose to monitor the model's relevancy signal and manipulate it such that the model is focused on the foreground object. This is done as a finetuning step, involving relatively few samples consisting of pairs of images and their associated foreground masks. Specifically, we encourage the model's relevancy map (i) to assign lower relevance to background regions, (ii) to consider as much information as possible from the foreground, and (iii) we encourage the decisions to have high confidence. When applied to Vision Transformer (ViT) models, a marked improvement in robustness to domain-shifts is observed. Moreover, the foreground masks can be obtained automatically, from a self-supervised variant of the ViT model itself; therefore no additional supervision is required.

## SAMURAI: Shape And Material from Unconstrained Real-world Arbitrary Image collections

- Mark Boss · Andreas Engelhardt · Abhishek Kar · Yuanzhen Li · Deqing Sun · Jonathan Barron · Hendrik PA Lensch · Varun Jampani
- abstract@[open-review](#): Inverse rendering of an object under entirely unknown capture conditions is a fundamental challenge in computer vision and graphics. Neural approaches such as NeRF have achieved photorealistic results on novel view synthesis, but they require known camera poses. Solving this problem with unknown camera poses is highly challenging as it requires joint optimization over shape, radiance, and pose. This problem is exacerbated when the input images are captured in the wild with varying backgrounds and illuminations. In such image collections in the wild, standard pose estimation techniques fail due to very few estimated correspondences across images. Furthermore, NeRF cannot relight a scene under any illumination, as it operates on radiance (the product of reflectance and illumination). We propose a joint optimization framework to estimate the shape, BRDF, and per-image camera pose and illumination. Our method works on in-the-wild online image collections of an object and produces relightable 3D assets for several use-cases such as AR/VR. To our knowledge, our method is the first to tackle this severely unconstrained task with minimal user interaction.

## Trading off Image Quality for Robustness is not Necessary with Deterministic Autoencoders

- Amrutha Saseendran · Kathrin Skubch · Margret Keuper
- abstract@[open-review](#): The susceptibility of Variational Autoencoders (VAEs) to adversarial attacks indicates the necessity to evaluate the robustness of the learned representations along with the generation performance. The vulnerability of VAEs has been attributed to the limitations associated with their variational formulation. Deterministic autoencoders could overcome the practical limitations associated with VAEs and offer a promising alternative for image generation applications. In this work, we propose an adversarially robust deterministic autoencoder with superior performance in terms of both generation and robustness of the learned representations. We introduce a regularization scheme to incorporate adversarially perturbed data points to the training pipeline without increasing the computational complexity or compromising the generation fidelity by leveraging a loss based on the two-point Kolmogorov–Smirnov test between representations. We conduct extensive experimental studies on popular image benchmark datasets to quantify the robustness of the proposed approach based on the adversarial attacks targeted at VAEs. Our empirical findings show that the proposed method achieves significant performance in both robustness and fidelity when compared to the robust VAE models.

## MonoSDF: Exploring Monocular Geometric Cues for Neural Implicit Surface Reconstruction

- Zehao Yu · Songyou Peng · Michael Niemeyer · Torsten Sattler · Andreas Geiger
- abstract@[open-review](#): In recent years, neural implicit surface reconstruction methods have become popular for multi-view 3D reconstruction. In contrast to traditional multi-view stereo methods, these approaches tend to produce smoother and more complete reconstructions due to the inductive smoothness bias of neural networks. State-of-the-art neural implicit methods allow for high-quality reconstructions of simple scenes from many input views. Yet, their performance drops significantly for larger and more complex scenes and scenes captured from sparse viewpoints. This is caused primarily by the inherent ambiguity in the RGB reconstruction loss that does not provide enough constraints, in particular in less-observed and textureless areas. Motivated by recent advances in the area of monocular geometry prediction, we systematically explore the utility these cues provide for improving neural implicit surface reconstruction. We demonstrate that depth and normal cues, predicted by general-purpose monocular estimators, significantly improve reconstruction quality and optimization time. Further, we analyse and investigate multiple design choices for representing neural implicit surfaces, ranging from monolithic MLP models over single-grid to multi-resolution grid representations. We observe that geometric monocular priors improve performance both for small-scale single-object as well as large-scale multi-object scenes, independent of the choice of representation.

## Online PAC-Bayes Learning

- Maxime Haddouche · Benjamin Guedj
- abstract@[open-review](#): Most PAC-Bayesian bounds hold in the batch learning setting where data is collected at once, prior to inference or prediction. This somewhat departs from many contemporary learning problems where data streams are collected and the algorithms must dynamically adjust. We prove new PAC-Bayesian bounds in this online learning framework, leveraging an updated definition of regret, and we revisit classical PAC-Bayesian results

with a batch-to-online conversion, extending their remit to the case of dependent data. Our results hold for bounded losses, potentially non-convex}, paving the way to promising developments in online learning.

## [Temporal Effective Batch Normalization in Spiking Neural Networks](#)

- Chaoteng Duan · Jianhao Ding · Shiyan Chen · Zhaofei Yu · Tiejun Huang
- abstract@[open-review](#): Spiking Neural Networks (SNNs) are promising in neuromorphic hardware owing to utilizing spatio-temporal information and sparse event-driven signal processing. However, it is challenging to train SNNs due to the non-differentiable nature of the binary firing function. The surrogate gradients alleviate the training problem and make SNNs obtain comparable performance as Artificial Neural Networks (ANNs) with the same structure. Unfortunately, batch normalization, contributing to the success of ANNs, does not play a prominent role in SNNs because of the additional temporal dimension. To this end, we propose an effective normalization method called temporal effective batch normalization (TEBN). By rescaling the presynaptic inputs with different weights at every time-step, temporal distributions become smoother and uniform. Theoretical analysis shows that TEBN can be viewed as a smoother of SNN's optimization landscape and could help stabilize the gradient norm. Experimental results on both static and neuromorphic datasets show that SNNs with TEBN outperform the state-of-the-art accuracy with fewer time-steps, and achieve better robustness to hyper-parameters than other normalizations.

## [Dynamic Inverse Reinforcement Learning for Characterizing Animal Behavior](#)

- Zoe Ashwood · Aditi Jha · Jonathan Pillow
- abstract@[open-review](#): Building computational models of decision-making is a core objective in both neuroscience and psychology. While many models have been developed for characterizing behavior in binary decision-making and bandit tasks, limited work has focused on animal decision-making in more complex tasks, such as navigation through a maze. Inverse reinforcement learning (IRL) is a promising direction for understanding such behavior as it aims to infer the unknown reward function of an agent from its trajectories. However, IRL has yet to be widely applied in neuroscience. One potential reason for this is that existing IRL frameworks assume that an agent's reward function is fixed over time. In this work we introduce 'DIRL', a novel IRL framework that allows for time-varying intrinsic rewards. Our method decomposes the unknown reward function into a linear combination of reward maps ('goal maps'), which can be weighted differently at each moment in time. We develop an inference method that allows us to recover these rewards, and demonstrate the application of our method in simulation, as well as on the trajectories of mice exploring a labyrinth. Our method returns interpretable reward functions for two separate cohorts of mice, and provides a novel characterization of exploratory behavior. Overall, we anticipate our framework having broad applicability in neuroscience, and in facilitating the design of biologically-inspired reward functions for training artificial agents to perform analogous tasks.

## [Stability and Generalization Analysis of Gradient Methods for Shallow Neural Networks](#)

- Yunwen Lei · Rong Jin · Yiming Ying
- abstract@[open-review](#): While significant theoretical progress has been achieved, unveiling the generalization mystery of overparameterized neural networks still remains largely elusive. In this paper, we study the generalization behavior of shallow neural networks (SNNs) by leveraging the concept of algorithmic stability. We consider gradient descent (GD) and stochastic gradient descent (SGD) to train SNNs, for both of which we develop consistent excess risk bounds by balancing the optimization and generalization via early-stopping. As compared to existing analysis on GD, our new analysis requires a relaxed overparameterization assumption and also applies to SGD. The key for the improvement is a better estimation of the smallest eigenvalues of the Hessian matrices of the empirical risks and the loss function along the trajectories of GD and SGD by providing a refined estimation of their iterates.

## [Cost-Sensitive Self-Training for Optimizing Non-Decomposable Metrics](#)

- Harsh Rangwani · shrinivas ramasubramanian · Sho Takemori · Kato Takashi · Yuhei Umeda · Venkatesh Babu R
- abstract@[open-review](#): Self-training based semi-supervised learning algorithms have enabled the learning of highly accurate deep neural networks, using only a fraction of labeled data. However, majority of work on self-training has focused on the objective of improving the accuracy whereas practical machine learning systems can have complex goals i.e. maximizing the minimum of recall across classes that are non-decomposable. In this work, we introduce the Cost-Sensitive Self-Training (CSST) framework which generalizes the self-training based methods for optimizing non-decomposable metrics. We prove that our framework is able to better optimize the desired non-decomposable metric, under similar data distribution assumptions made for the analysis of self-training. Using the proposed CSST framework we obtain practical self-training methods (for both vision and NLP tasks) for optimizing different non-decomposable metrics using deep neural networks. Our results demonstrate that CSST achieves an improvement over the state-of-the-art in the majority of cases.

## [Censored Quantile Regression Neural Networks](#)

- Tim Pearce · Jong-Hyeon Jeong · yichen jia · Jun Zhu
- abstract@[open-review](#): This paper considers doing quantile regression on censored data using neural networks (NNs). This adds to the survival analysis toolkit by allowing direct prediction of the target variable, along with a distribution-free characterisation of uncertainty, using a flexible function approximator. We begin by showing how an algorithm popular in linear models can be applied to NNs. However, the resulting procedure is inefficient, requiring sequential optimisation of an individual NN at each desired quantile. Our major contribution is a novel algorithm that simultaneously optimises a grid of quantiles output by a single NN. To offer theoretical insight into our algorithm, we show firstly that it can be interpreted as a form of expectation-maximisation, and secondly that it exhibits a desirable 'self-correcting' property. Experimentally, the algorithm produces quantiles that are better calibrated than existing methods on 10 out of 12 real datasets.

## [Keypoint-Guided Optimal Transport with Applications in Heterogeneous Domain Adaptation](#)

- Xiang Gu · Yucheng Yang · Wei Zeng · Jian Sun · Zongben Xu
- abstract@[open-review](#): Existing Optimal Transport (OT) methods mainly derive the optimal transport plan/matching under the criterion of transport cost/distance minimization, which may cause incorrect matching in some cases. In many applications, annotating a few matched keypoints across domains is reasonable or even effortless in annotation burden. It is valuable to investigate how to leverage the annotated keypoints to guide the correct matching in OT. In this paper, we propose a novel KeyPoint-Guided model by ReLation preservation (KPG-RL) that searches for the matching guided by the keypoints in OT. To impose the keypoints in OT, first, we propose a mask-based constraint of the transport plan that preserves the matching of keypoint pairs. Second, we propose to preserve the relation of each data point to the keypoints to guide the matching. The proposed KPG-RL model can be solved by the Sinkhorn's algorithm and is applicable even when distributions are supported in different spaces. We further utilize the relation preservation constraint in the Kantorovich Problem and Gromov-Wasserstein model to impose the guidance of keypoints in them. Meanwhile, the proposed KPG-RL model is extended to partial OT setting. As an application, we apply the proposed KPG-RL model to the heterogeneous domain adaptation. Experiments verified the effectiveness of the KPG-RL model.

## [TUSK: Task-Agnostic Unsupervised Keypoints](#)

- Yuhe Jin · Weiwei Sun · Jan Hosang · Eduard Trulls · Kwang Moo Yi
- abstract@[open-review](#): Existing unsupervised methods for keypoint learning rely heavily on the assumption that a specific keypoint type (e.g. elbow, digit, abstract geometric shape) appears only once in an image. This greatly limits their applicability, as each instance must be isolated before applying the method—an issue that is never discussed or evaluated. We thus propose a novel method to learn Task-agnostic, UnSupervised Keypoints (TUSK) which can deal with multiple instances. To achieve this, instead of the commonly-used strategy of detecting multiple heatmaps, each dedicated to a specific keypoint type, we use a single heatmap for detection, and enable unsupervised learning of keypoint types through clustering. Specifically, we encode semantics into the keypoints by teaching them to reconstruct images from a sparse set of keypoints and their descriptors, where the descriptors are forced to form distinct clusters in feature space around learned prototypes. This makes our approach amenable to a wider range of tasks than any previous unsupervised keypoint method: we show experiments on multiple-instance detection and classification, object discovery, and landmark detection—all unsupervised—with performance on par with the state of the art, while also being able to deal with multiple instances.

## [Dataset Inference for Self-Supervised Models](#)

- Adam Dziedzic · Haonan Duan · Muhammad Ahmad Kaleem · Nikita Dhawan · Jonas Guan · Yannis Cattan · Franziska Boenisch · Nicolas Papernot
- abstract@[open-review](#): Self-supervised models are increasingly prevalent in machine learning (ML) since they reduce the need for expensively labeled data. Because of their versatility in downstream applications, they are increasingly used as a service exposed via public APIs. At the same time, these encoder models are particularly vulnerable to model stealing attacks due to the high dimensionality of vector representations they output. Yet, encoders remain undefended: existing mitigation strategies for stealing attacks focus on supervised learning. We introduce a new dataset inference defense, which uses the private training set of the victim encoder model to attribute its ownership in the event of stealing. The intuition is that the log-likelihood of an encoder's output representations is higher on the victim's training data than on test data if it is stolen from the victim, but not if it is independently trained. We compute this log-likelihood using density estimation models. As part of our evaluation, we also propose measuring the fidelity of stolen encoders and quantifying the effectiveness of the theft detection without involving downstream tasks; instead, we leverage mutual information and distance measurements. Our extensive empirical results in the vision domain demonstrate that dataset inference is a promising direction for defending self-supervised models against model stealing.

## [SNAKE: Shape-aware Neural 3D Keypoint Field](#)

- Chengliang Zhong · Peixing You · Xiaoxue Chen · Hao Zhao · Fuchun Sun · Guyue Zhou · Xiaodong Mu · Chuang Gan · Wenbing Huang
- abstract@[open-review](#): Detecting 3D keypoints from point clouds is important for shape reconstruction, while this work investigates the dual question: can shape reconstruction benefit 3D keypoint detection? Existing methods either seek salient features according to statistics of different orders or learn to predict keypoints that are invariant to transformation. Nevertheless, the idea of incorporating shape reconstruction into 3D keypoint detection is under-explored. We argue that this is restricted by former problem formulations. To this end, a novel unsupervised paradigm named SNAKE is proposed, which is short for shape-aware neural 3D keypoint field. Similar to recent coordinate-based radiance or distance field, our network takes 3D coordinates as inputs and predicts implicit shape indicators and keypoint saliency simultaneously, thus naturally entangling 3D keypoint detection and shape reconstruction. We achieve superior performance on various public benchmarks, including standalone object datasets ModelNet40, KeypointNet, SMPL meshes and scene-level datasets 3DMatch and Redwood. Intrinsic shape awareness brings several advantages as follows. (1) SNAKE generates 3D keypoints consistent with human semantic annotation, even without such supervision. (2) SNAKE outperforms counterparts in terms of repeatability, especially when the input point clouds are down-sampled. (3) the generated keypoints allow accurate geometric registration, notably in a zero-shot setting. Codes and models will be released.

## [Formalizing Coherence and Consistency Applied to Transfer Learning in Neuro-Symbolic Autoencoders](#)

- Harald Stromfelt · Luke Dickens · Artur Garcez · Alessandra Russo
- abstract@[open-review](#): In the study of reasoning in neural networks, recent efforts have sought to improve coherence and consistency of neural sequence models. This is an important development in the study of neuro-symbolic systems. In symbolic AI, however, the concepts of consistency and coherence are defined formally. The provision of such formal definitions is needed to offer a common basis for the quantitative evaluation and systematic comparison of connectionist, neuro-symbolic and transfer learning approaches. In this paper we introduce formal definitions for coherence and consistency of neural systems. To illustrate the usefulness of the definitions, we propose a new dynamic relation-decoder model built around the principles of consistency and coherence. By comparing several existing relation-decoders on a partial relation transfer learning task and novel data set introduced in this paper, our experiments show that relation-decoders that can maintain consistency over unobserved regions of representation space, retain coherence across domains and achieve better transfer learning performance.

## [Dynamic Graph Neural Networks Under Spatio-Temporal Distribution Shift](#)

- Zeyang Zhang · Xin Wang · Ziwei Zhang · Haoyang Li · Zhou Qin · Wenwu Zhu
- abstract@[open-review](#): Dynamic graph neural networks (DyGNNs) have demonstrated powerful predictive abilities by exploiting graph structural and temporal dynamics. However, the existing DyGNNs fail to handle distribution shifts, which naturally exist in dynamic graphs, mainly because the patterns exploited by DyGNNs may be variant with respect to labels under distribution shifts. In this paper, we propose to handle spatio-temporal distribution shifts in dynamic graphs by discovering and utilizing {it invariant patterns}, i.e., structures and features whose predictive abilities are stable across distribution shifts, which faces two key challenges: 1) How to discover the complex variant and invariant spatio-temporal patterns in dynamic graphs, which involve both time-varying graph structures and node features. 2) How to handle spatio-temporal distribution shifts with the discovered variant and invariant patterns. To tackle these challenges, we propose the Disentangled Intervention-based Dynamic graph Attention networks (DIDA). Our proposed method can effectively handle spatio-temporal distribution shifts in dynamic graphs by discovering and fully utilizing invariant spatio-temporal patterns. Specifically, we first propose a disentangled spatio-temporal attention network to capture the variant and invariant patterns. Then, we design a spatio-temporal intervention mechanism to create multiple interventional distributions by sampling and reassembling variant patterns across neighborhoods and time stamps to eliminate the spurious impacts of variant patterns. Lastly, we propose an invariance regularization term to minimize the variance of predictions in intervened distributions so that our model can make predictions based on invariant patterns with stable predictive abilities and therefore handle distribution shifts. Experiments on three real-world datasets and one synthetic dataset demonstrate the superiority of our method over state-of-the-art baselines under distribution shifts. Our work is the first study of DyGNNs under distribution shifts, to the best of our knowledge.

## [ComGAN: Unsupervised Disentanglement and Segmentation via Image Composition](#)

- Rui Ding · Kehua Guo · Xiangyuan Zhu · Zheng Wu · Liwei Wang
- abstract@[open-review](#): We propose ComGAN, a simple unsupervised generative model, which simultaneously generates realistic images and high semantic masks under an adversarial loss and a binary regularization. In this paper, we first investigate two kinds of trivial solutions in the compositional generation process, and demonstrate their source is vanishing gradients on the mask. Then, we solve trivial solutions from the perspective of architecture. Furthermore, we redesign two fully unsupervised modules based on ComGAN (DS-ComGAN), where the {d}isentanglement module associates the foreground, background and mask with three independent variables, and the {s}egmentation module learns object segmentation. Experimental results show that (i) ComGAN's network architecture effectively avoids trivial solutions without any supervised information and regularization; (ii) DS-ComGAN achieves remarkable results and outperforms existing semi-supervised and weakly supervised methods by a large margin in both the image disentanglement and unsupervised segmentation tasks. It implies that the redesign of ComGAN is a possible direction for future unsupervised work.

## Pure Transformers are Powerful Graph Learners

- Jinwoo Kim · Dat Nguyen · Seonwoo Min · Sungjun Cho · Moontae Lee · Honglak Lee · Seunghoon Hong
- abstract@[open-review](#): We show that standard Transformers without graph-specific modifications can lead to promising results in graph learning both in theory and practice. Given a graph, we simply treat all nodes and edges as independent tokens, augment them with token embeddings, and feed them to a Transformer. With an appropriate choice of token embeddings, we prove that this approach is theoretically at least as expressive as an invariant graph network (2-IGN) composed of equivariant linear layers, which is already more expressive than all message-passing Graph Neural Networks (GNN). When trained on a large-scale graph dataset (PCQM4Mv2), our method coined Soft Graph Transformer (SGT) achieves significantly better results compared to GNN baselines and competitive results compared to Transformer variants with sophisticated graph-specific inductive bias.

## Gaussian Copula Embeddings

- Chien Lu · Jaakko Peltonen
- abstract@[open-review](#): Learning latent vector representations via embedding models has been shown promising in machine learning. However, most of the embedding models are still limited to a single type of observation data. We propose a Gaussian copula embedding model to learn latent vector representations of items in a heterogeneous data setting. The proposed model can effectively incorporate different types of observed data and, at the same time, yield robust embeddings. We demonstrate the proposed model can effectively learn in many different scenarios, outperforming competing models in modeling quality and task performance.

## [PatchComplete: Learning Multi-Resolution Patch Priors for 3D Shape Completion on Unseen Categories](#)

- Yuchen Rao · Yinyu Nie · Angela Dai
- abstract@[open-review](#): While 3D shape representations enable powerful reasoning in many visual and perception applications, learning 3D shape priors tends to be constrained to the specific categories trained on, leading to an inefficient learning process, particularly for general applications with unseen categories. Thus, we propose PatchComplete, which learns effective shape priors based on multi-resolution local patches, which are often more general than full shapes (e.g., chairs and tables often both share legs) and thus enable geometric reasoning about unseen class categories. To learn these shared substructures, we learn multi-resolution patch priors across all train categories, which are then associated to input partial shape observations by attention across the patch priors, and finally decoded into a complete shape reconstruction. Such patch-based priors avoid overfitting to specific train categories and enable reconstruction on entirely unseen categories at test time. We demonstrate the effectiveness of our approach on synthetic ShapeNet data as well as challenging real-scanned objects from ScanNet, which include noise and clutter, improving over state of the art in novel-category shape completion by 19.3% in chamfer distance on ShapeNet, and 9.0% for ScanNet.

## [GLIF: A Unified Gated Leaky Integrate-and-Fire Neuron for Spiking Neural Networks](#)

- Xingting Yao · Fanrong Li · Zitao Mo · Jian Cheng
- abstract@[open-review](#): Spiking Neural Networks (SNNs) have been studied over decades to incorporate their biological plausibility and leverage their promising energy efficiency. Throughout existing SNNs, the leaky integrate-and-fire (LIF) model is commonly adopted to formulate the spiking neuron and evolves into numerous variants with different biological features. However, most LIF-based neurons support only single biological feature in different neuronal behaviors, limiting their expressiveness and neuronal dynamic diversity. In this paper, we propose GLIF, a unified spiking neuron, to fuse different bio-features in different neuronal behaviors, enlarging the representation space of spiking neurons. In GLIF, gating factors, which are exploited to determine the proportion of the fused bio-features, are learnable during training. Combining all learnable membrane-related parameters, our method can make spiking neurons different and constantly changing, thus increasing the heterogeneity and adaptivity of spiking neurons. Extensive experiments on a variety of datasets demonstrate that our method obtains superior performance compared with other SNNs by simply changing their neuronal formulations to GLIF. In particular, we train a spiking ResNet-19 with GLIF and achieve \$77.35\%\$ top-1 accuracy with six time steps on CIFAR-100, which has advanced the state-of-the-art. Codes are available at <https://github.com/Ikarosy/Gated-LIF>.

## [BR-SNIS: Bias Reduced Self-Normalized Importance Sampling](#)

- Gabriel Cardoso · Sergey Samsonov · Achille Thin · Eric Moulines · Jimmy Olsson
- abstract@[open-review](#): Importance Sampling (IS) is a method for approximating expectations with respect to a target distribution using independent samples from a proposal distribution and the associated to importance weights. In many cases, the target distribution is known up to a normalization constant and self-normalized IS (SNIS) is then used. While the use of self-normalization can have a positive effect on the dispersion of the estimator, it introduces bias. In this work, we propose a new method BR-SNIS whose complexity is essentially the same as SNIS and which significantly reduces bias. This method is a wrapper, in the sense that it uses the same proposal samples and importance weights but makes a clever use of iterated sampling-importance-resampling (i-SIR) to form a bias-reduced version of the estimator. We derive the proposed algorithm with rigorous theoretical results, including novel bias, variance, and high-probability bounds. We illustrate our findings with numerical examples.

## [Shape, Light, and Material Decomposition from Images using Monte Carlo Rendering and Denoising](#)

- Jon Hasselgren · Nikolai Hofmann · Jacob Munkberg
- abstract@[open-review](#): Recent advances in differentiable rendering have enabled high-quality reconstruction of 3D scenes from multi-view images. Most methods rely on simple rendering algorithms: pre-filtered direct lighting or learned representations of irradiance. We show that a more realistic shading model, incorporating ray tracing and Monte Carlo integration, substantially improves decomposition into shape, materials & lighting. Unfortunately, Monte Carlo integration provides estimates with significant noise, even at large sample counts, which makes gradient-based inverse rendering very challenging. To address this, we incorporate multiple importance sampling and denoising in a novel inverse rendering pipeline. This improves convergence and enables gradient-based optimization at low sample counts. We present an efficient method to jointly reconstruct geometry (explicit triangle meshes), materials, and lighting, which substantially improves material and light separation compared to previous work. We argue that denoising can become an integral part of high quality inverse rendering pipelines.

## [TTOpt: A Maximum Volume Quantized Tensor Train-based Optimization and its Application to Reinforcement Learning](#)

- Konstantin Sozykin · Andrei Chertkov · Roman Schutski · Anh-Huy Phan · Andrzej S CICHOCKI · Ivan Oseledets
- abstract@[open-review](#): We present a novel procedure for optimization based on the combination of efficient quantized tensor train representation and a generalized maximum matrix volume principle. We demonstrate the applicability of the new Tensor Train Optimizer (TTOpt) method for various tasks, ranging from minimization of multidimensional functions to reinforcement learning. Our algorithm compares favorably to popular gradient-free methods and outperforms them by the number of function evaluations or execution time, often by a significant margin.

## [Invariance-Aware Randomized Smoothing Certificates](#)

- Jan Schuchardt · Stephan Gähnemann

- abstract@[open-review](#): Building models that comply with the invariances inherent to different domains, such as invariance under translation or rotation, is a key aspect of applying machine learning to real world problems like molecular property prediction, medical imaging, protein folding or LiDAR classification. For the first time, we study how the invariances of a model can be leveraged to provably guarantee the robustness of its predictions. We propose a gray-box approach, enhancing the powerful black-box randomized smoothing technique with white-box knowledge about invariances. First, we develop a post-processing-based gray-box certification procedure that can be applied to arbitrary models with invariance under permutation and Euclidean isometries. Then, we derive provably tight gray-box certificates. We experimentally demonstrate that the provably tight certificates can offer much stronger guarantees, but that in practical scenarios the post-processing method is a good approximation.

## [Contextual Squeeze-and-Excitation for Efficient Few-Shot Image Classification](#)

- Massimiliano Patacchiola · John Bronskill · Aliaksandra Shysheya · Katja Hofmann · Sebastian Nowozin · Richard Turner
- abstract@[open-review](#): Recent years have seen a growth in user-centric applications that require effective knowledge transfer across tasks in the low-data regime. An example is personalization, where a pretrained system is adapted by learning on small amounts of labeled data belonging to a specific user. This setting requires high accuracy under low computational complexity, therefore the Pareto frontier of accuracy vs. adaptation cost plays a crucial role. In this paper we push this Pareto frontier in the few-shot image classification setting with a key contribution: a new adaptive block called Contextual Squeeze-and-Excitation (CaSE) that adjusts a pretrained neural network on a new task to significantly improve performance with a single forward pass of the user data (context). We use meta-trained CaSE blocks to conditionally adapt the body of a network and a fine-tuning routine to adapt a linear head, defining a method called UpperCaSE. UpperCaSE achieves a new state-of-the-art accuracy relative to meta-learners on the 26 datasets of VTAB+MD and on a challenging real-world personalization benchmark (ORBIT), narrowing the gap with leading fine-tuning methods with the benefit of orders of magnitude lower adaptation cost.

## [Rule-Based but Flexible? Evaluating and Improving Language Models as Accounts of Human Moral Judgment](#)

- Zhijing Jin · Sydney Levine · Fernando Gonzalez Adauto · Ojasv Kamal · Maarten Sap · Mrinmaya Sachan · Rada Mihalcea · Josh Tenenbaum · Bernhard Schölkopf
- abstract@[open-review](#): AI systems are becoming increasingly intertwined with human life. In order to effectively collaborate with humans and ensure safety, AI systems need to be able to understand, interpret and predict human moral judgments and decisions. Human moral judgments are often guided by rules, but not always. A central challenge for AI safety is capturing the flexibility of the human moral mind — the ability to determine when a rule should be broken, especially in novel or unusual situations. In this paper, we present a novel challenge set consisting of rule-breaking question answering (RBQA) of cases that involve potentially permissible rule-breaking -- inspired by recent moral psychology studies. Using a state-of-the-art large language model (LLM) as a basis, we propose a novel moral chain of thought (MoralCoT) prompting strategy that combines the strengths of LLMs with theories of moral reasoning developed in cognitive science to predict human moral judgments. MoralCoT outperforms seven existing LLMs by 9.04% F1, suggesting that modeling human reasoning might be necessary to capture the flexibility of the human moral mind. We also conduct a detailed error analysis to suggest directions for future work to improve AI safety using RBQA.

## [VaiPhy: a Variational Inference Based Algorithm for Phylogeny](#)

- Hazal Koptagel · Oskar Kviman · Harald Melin · Negar Safinianaini · Jens Lagergren
- abstract@[open-review](#): Phylogenetics is a classical methodology in computational biology that today has become highly relevant for medical investigation of single-cell data, e.g., in the context of development of cancer. The exponential size of the tree space is unfortunately a formidable obstacle for current Bayesian phylogenetic inference using Markov chain Monte Carlo based methods since these rely on local operations. And although more recent variational inference (VI) based methods offer speed improvements, they rely on expensive auto-differentiation operations for learning the variational parameters. We propose VaiPhy, a remarkably fast VI based algorithm for approximate posterior inference in an \textit{augmented tree space}. VaiPhy produces marginal log-likelihood estimates on par with the state-of-the-art methods on real data, and is considerably faster since it does not require auto-differentiation. Instead, VaiPhy combines coordinate ascent update equations with two novel sampling schemes: (i) \textit{SLANTIS}, a proposal distribution for tree topologies in the augmented tree space, and (ii) the \textit{JC sampler}, the, to the best of our knowledge, first ever scheme for sampling branch lengths directly from the popular Jukes-Cantor model. We compare VaiPhy in terms of density estimation and runtime. Additionally, we evaluate the reproducibility of the baselines. We provide our code on GitHub: \url{gitlink}.

## [CATER: Intellectual Property Protection on Text Generation APIs via Conditional Watermarks](#)

- Xuanli He · Qiongkai Xu · Yi Zeng · Lingjuan Lyu · Fangzhao Wu · Jiwei Li · Ruoxi Jia
- abstract@[open-review](#): Previous works have validated that text generation APIs can be stolen through imitation attacks, causing IP violations. In order to protect the IP of text generation APIs, a recent work has introduced a watermarking algorithm and utilized the null-hypothesis test as a post-hoc ownership verification on the imitation models. However, we find that it is possible to detect those watermarks via sufficient statistics of the frequencies of candidate watermarking words. To address this drawback, in this paper, we propose a novel Conditional wATERmarking framework (CATER) for protecting the IP of text generation APIs. An optimization method is proposed to decide the watermarking rules that can minimize the distortion of overall word distributions while maximizing the change of conditional word selections. Theoretically, we prove that it is infeasible for even the savviest attacker (they know how CATER works) to reveal the used watermarks from a large pool of potential word pairs based on statistical inspection. Empirically, we observe that high-order conditions lead to an exponential growth of suspicious (unused) watermarks, making our crafted watermarks more stealthy. In addition, CATER can effectively identify the IP infringement under architectural mismatch and cross-domain imitation attacks, with negligible impairments on the generation quality of victim APIs. We envision our work as a milestone for stealthily protecting the IP of text generation APIs.

## [Obj2Seq: Formatting Objects as Sequences with Class Prompt for Visual Tasks](#)

- Zhiyang Chen · Yousong Zhu · Zhaowen Li · Fan Yang · Wei Li · Haixin Wang · Chaoyang Zhao · Liwei Wu · Rui Zhao · Jinqiao Wang · Ming Tang
- abstract@[open-review](#): Visual tasks vary a lot in their output formats and concerned contents, therefore it is hard to process them with an identical structure. One main obstacle lies in the high-dimensional outputs in object-level visual tasks. In this paper, we propose an object-centric vision framework, Obj2Seq. Obj2Seq takes objects as basic units, and regards most object-level visual tasks as sequence generation problems of objects. Therefore, these visual tasks can be decoupled into two steps. First recognize objects of given categories, and then generate a sequence for each of these objects. The definition of the output sequences varies for different tasks, and the model is supervised by matching these sequences with ground-truth targets. Obj2Seq is able to flexibly determine input categories to satisfy customized requirements, and be easily extended to different visual tasks. When experimenting on MS COCO, Obj2Seq achieves 45.7% AP on object detection, 89.0% AP on multi-label classification and 65.0% AP on human pose estimation. These results demonstrate its potential to be generally applied to different visual tasks. Code has been made available at: <https://github.com/CASIA-IVA-Lab/Obj2Seq>.

## [Low-rank lottery tickets: finding efficient low-rank neural networks via matrix differential equations](#)

- Steffen Schottenhamer · Emanuele Zangrandi · Jonas Kusch · Gianluca Ceruti · Francesco Tudisco
- abstract@[open-review](#): Neural networks have achieved tremendous success in a large variety of applications. However, their memory footprint and computational demand can render them impractical in application settings with limited hardware or energy resources. In this work, we propose a novel

algorithm to find efficient low-rank subnetworks. Remarkably, these subnetworks are determined and adapted already during the training phase and the overall time and memory resources required by both training and evaluating them is significantly reduced. The main idea is to restrict the weight matrices to a low-rank manifold and to update the low-rank factors rather than the full matrix during training. To derive training updates that are restricted to the prescribed manifold, we employ techniques from dynamic model order reduction for matrix differential equations. Moreover, our method automatically and dynamically adapts the ranks during training to achieve a desired approximation accuracy. The efficiency of the proposed method is demonstrated through a variety of numerical experiments on fully-connected and convolutional networks.

## [Curriculum Reinforcement Learning using Optimal Transport via Gradual Domain Adaptation](#)

- Peide Huang · Mengdi Xu · Jiacheng Zhu · Laixi Shi · Fei Fang · DING ZHAO
- abstract@[open-review](#): Curriculum Reinforcement Learning (CRL) aims to create a sequence of tasks, starting from easy ones and gradually learning towards some difficult tasks. In this work, we focus on the idea of framing CRL as interpolations between a source (auxiliary) and a target task distribution. Although existing studies have shown the great potential of this idea, it remains unclear how to formally quantify and generate the movement between task distributions. Inspired by the insights from gradual domain adaptation in semi-supervised learning, we create a natural curriculum by breaking down the potentially large task distributional shift in CRL into smaller shifts. We propose GRADIENT, which formulates CRL as an optimal transport problem with a tailored distance metric between tasks. Specifically, we generate a sequence of task distributions as a geodesic interpolation between the source and target distributions, i.e., the Wasserstein barycenters. Different from many existing methods, our algorithm considers a task-dependent contextual distance metric and is capable of handling non-parametric distributions in both continuous and discrete context settings. In addition, we theoretically show that GRADIENT enables smooth transferring between subsequent stages in the curriculum under certain conditions. Our empirical results demonstrate that the proposed algorithm achieves high performance in terms of learning efficiency and asymptotic performance in a wide range of tasks.

## [Causal Discovery in Heterogeneous Environments Under the Sparse Mechanism Shift Hypothesis](#)

- Ronan Perry · Julius von Kägelgen · Bernhard Schölkopf
- abstract@[open-review](#): Machine learning approaches commonly rely on the assumption of independent and identically distributed (i.i.d.) data. In reality, however, this assumption is almost always violated due to distribution shifts between environments. Although valuable learning signals can be provided by heterogeneous data from changing distributions, it is also known that learning under arbitrary (adversarial) changes is impossible. Causality provides a useful framework for modeling distribution shifts, since causal models encode both observational and interventional distributions. In this work, we explore the sparse mechanism shift hypothesis which posits that distribution shifts occur due to a small number of changing causal conditionals. Motivated by this idea, we apply it to learning causal structure from heterogeneous environments, where i.i.d. data only allows for learning an equivalence class of graphs without restrictive assumptions. We propose the Mechanism Shift Score (MSS), a score-based approach amenable to various empirical estimators, which provably identifies the entire causal structure with high probability if the sparse mechanism shifts hypothesis holds. Empirically, we verify behavior predicted by the theory and compare multiple estimators and score functions to identify the best approaches in practice. Compared to other methods, we show how MSS bridges a gap by both being nonparametric as well as explicitly leveraging sparse changes.

## [A Direct Approximation of AIXI Using Logical State Abstractions](#)

- Samuel Yang-Zhao · Tianyu Wang · Kee Siong Ng
- abstract@[open-review](#): We propose a practical integration of logical state abstraction with AIXI, a Bayesian optimality notion for reinforcement learning agents, to significantly expand the model class that AIXI agents can be approximated over to complex history-dependent and structured environments. The state representation and reasoning framework is based on higher-order logic, which can be used to define and enumerate complex features on non-Markovian and structured environments. We address the problem of selecting the right subset of features to form state abstractions by adapting the \$\Phi\$-MDP optimisation criterion from state abstraction theory. Exact Bayesian model learning is then achieved using a suitable generalisation of Context Tree Weighting over abstract state sequences. The resultant architecture can be integrated with different planning algorithms. Experimental results on controlling epidemics on large-scale contact networks validates the agent's performance.

## [The Policy-gradient Placement and Generative Routing Neural Networks for Chip Design](#)

- Ruoyu Cheng · Xianglong Lyu · Yang Li · Junjie Ye · Jianye Hao · Junchi Yan
- abstract@[open-review](#): Placement and routing are two critical yet time-consuming steps of chip design in modern VLSI systems. Distinct from traditional heuristic solvers, this paper on one hand proposes an RL-based model for mixed-size macro placement, which differs from existing learning-based placers that often consider the macro by coarse grid-based mask. While the standard cells are placed via gradient-based GPU acceleration. On the other hand, a one-shot conditional generative routing model, which is composed of a special-designed input-size-adapting generator and a bi-discriminator, is devised to perform one-shot routing to the pins within each net, and the order of nets to route is adaptively learned. Combining these techniques, we develop a flexible and efficient neural pipeline, which to our best knowledge, is the first joint placement and routing network without involving any traditional heuristic solver. Experimental results on chip design benchmarks showcase the effectiveness of our approach, with code that will be made publicly available.

## [Reinforcement Learning with a Terminator](#)

- Guy Tennenholtz · Nadav Merlis · Lior Shani · Shie Mannor · Uri Shalit · Gal Chechik · Assaf Hallak · Gal Dalal
- abstract@[open-review](#): We present the problem of reinforcement learning with exogenous termination. We define the Termination Markov Decision Process (TerMDP), an extension of the MDP framework, in which episodes may be interrupted by an external non-Markovian observer. This formulation accounts for numerous real-world situations, such as a human interrupting an autonomous driving agent for reasons of discomfort. We learn the parameters of the TerMDP and leverage the structure of the estimation problem to provide state-wise confidence bounds. We use these to construct a provably-efficient algorithm, which accounts for termination, and bound its regret. Motivated by our theoretical analysis, we design and implement a scalable approach, which combines optimism (w.r.t. termination) and a dynamic discount factor, incorporating the termination probability. We deploy our method on high-dimensional driving and MinAtar benchmarks. Additionally, we test our approach on human data in a driving setting. Our results demonstrate fast convergence and significant improvement over various baseline approaches.

## [D^2NeRF: Self-Supervised Decoupling of Dynamic and Static Objects from a Monocular Video](#)

- Tianhao Wu · Fangcheng Zhong · Andrea Tagliasacchi · Forrester Cole · Cengiz Oztireli
- abstract@[open-review](#): Given a monocular video, segmenting and decoupling dynamic objects while recovering the static environment is a widely studied problem in machine intelligence. Existing solutions usually approach this problem in the image domain, limiting their performance and understanding of the environment. We introduce Decoupled Dynamic Neural Radiance Field (D^2NeRF), a self-supervised approach that takes a monocular video and learns a 3D scene representation which decouples moving objects, including their shadows, from the static background. Our method represents the moving objects and the static background by two separate neural radiance fields with only one allowing for temporal changes. A naive implementation of this approach leads to the dynamic component taking over the static one as the representation of the former is inherently more general and prone to overfitting. To this end, we propose a novel loss to promote correct separation of phenomena. We further propose a shadow field network to detect and decouple dynamically moving shadows. We introduce a new dataset containing various dynamic objects and shadows and demonstrate that our method can achieve

better performance than state-of-the-art approaches in decoupling dynamic and static 3D objects, occlusion and shadow removal, and image segmentation for moving objects.

## [EGSDE: Unpaired Image-to-Image Translation via Energy-Guided Stochastic Differential Equations](#)

- Min Zhao · Fan Bao · Chongxuan LI · Jun Zhu
- abstract@[open-review](#): Score-based diffusion generative models (SDGMs) have achieved the SOTA FID results in unpaired image-to-image translation (I2I). However, we notice that existing methods totally ignore the training data in the source domain, leading to sub-optimal solutions for unpaired I2I. To this end, we propose energy-guided stochastic differential equations (EGSDE) that employs an energy function pretrained on both the source and target domains to guide the inference process of a pretrained SDE for realistic and faithful unpaired I2I. Building upon two feature extractors, we carefully design the energy function such that it encourages the transferred image to preserve the domain-independent features and discard domain-specific ones. Further, we provide an alternative explanation of the EGSDE as a product of experts, where each of the three experts (corresponding to the SDE and two feature extractors) solely contributes to faithfulness or realism. Empirically, we compare EGSDE to a large family of baselines on three widely-adopted unpaired I2I tasks under four metrics. EGSDE not only consistently outperforms existing SDGMs-based methods in almost all settings but also achieves the SOTA realism results (e.g., FID of 65.82 in Cat †' Dog and FID of 59.75 in Wild †' Dog on AFHQ) without harming the faithful performance.

## [Alleviating ``Posterior Collapse'' in Deep Topic Models via Policy Gradient](#)

- Yewen Li · Chaojie Wang · Zhibin Duan · Dongsheng Wang · Bo Chen · Bo An · Mingyuan Zhou
- abstract@[open-review](#): Deep topic models have been proven as a promising way to extract hierarchical latent representations from documents represented as high-dimensional bag-of-words vectors. However, the representation capability of existing deep topic models is still limited by the phenomenon of "posterior collapse", which has been widely criticized in deep generative models, resulting in the higher-level latent representations exhibiting similar or meaningless patterns. To this end, in this paper, we first develop a novel deep-coupling generative process for existing deep topic models, which incorporates skip connections into the generation of documents, enforcing strong links between the document and its multi-layer latent representations. After that, utilizing data augmentation techniques, we reformulate the deep-coupling generative process as a Markov decision process and develop a corresponding Policy Gradient (PG) based training algorithm, which can further alleviate the information reduction at higher layers. Extensive experiments demonstrate that our developed methods can effectively alleviate "posterior collapse" in deep topic models, contributing to providing higher-quality latent document representations.

## [Out-of-Distribution Detection with An Adaptive Likelihood Ratio on Informative Hierarchical VAE](#)

- Yewen Li · Chaojie Wang · Xiaobo Xia · Tongliang Liu · xin miao · Bo An
- abstract@[open-review](#): Unsupervised out-of-distribution (OOD) detection is essential for the reliability of machine learning. In the literature, existing work has shown that higher-level semantics captured by hierarchical VAEs can be used to detect OOD instances. However, we empirically show that, the inherent issue of hierarchical VAEs, i.e., `posterior collapse', would seriously limit their capacity for OOD detection. Based on a thorough analysis for posterior collapse', we propose a novel informative hierarchical VAE to alleviate this issue through enhancing the connections between the data sample and its multi-layer stochastic latent representations during training. Furthermore, we propose a novel score function for unsupervised OOD detection, referred to as Adaptive Likelihood Ratio. With this score function, one can selectively aggregate the semantic information on multiple hidden layers of hierarchical VAEs, leading to a strong separability between in-distribution and OOD samples. Experimental results demonstrate that our method can significantly outperform existing state-of-the-art unsupervised OOD detection approaches.

## [Estimating the Arc Length of the Optimal ROC Curve and Lower Bounding the Maximal AUC](#)

- Song Liu
- abstract@[open-review](#): In this paper, we show the arc length of the optimal ROC curve is an  $\$f\$$ -divergence. By leveraging this result, we express the arc length using a variational objective and estimate it accurately using positive and negative samples. We show this estimator has a non-parametric convergence rate  $\$O_p(n^{-\beta/4})\$$  ( $\beta \in (0, 1]$  depends on the smoothness). Using the same technique, we show the surface area sandwiched between the optimal ROC curve and the diagonal can be expressed via a similar variational objective. These new insights lead to a novel two-step classification procedure that maximizes an approximate lower bound of the maximal AUC. Experiments on CIFAR-10 datasets show the proposed two-step procedure achieves good AUC performance in imbalanced binary classification tasks while being less computationally demanding than the classic AUC maximizer  $\text{AUC}$  in the offline setting.

## [IMED-RL: Regret optimal learning of ergodic Markov decision processes](#)

- Fabien Pesquerel · Odalric-Ambrym Maillard
- abstract@[open-review](#): We consider reinforcement learning in a discrete, undiscounted, infinite-horizon Markov decision problem (MDP) under the average reward criterion, and focus on the minimization of the regret with respect to an optimal policy, when the learner does not know the rewards nor transitions of the MDP. In light of their success at regret minimization in multi-armed bandits, popular bandit strategies, such as the optimistic UCB, KL-UCB or the Bayesian Thompson sampling strategy, have been extended to the MDP setup. Despite some key successes, existing strategies for solving this problem either fail to be provably asymptotically optimal, or suffer from prohibitive burn-in phase and computational complexity when implemented in practice. In this work, we shed a novel light on regret minimization strategies, by extending to reinforcement learning the computationally appealing Indexed Minimum Empirical Divergence (IMED) bandit algorithm. Traditional asymptotic problem-dependent lower bounds on the regret are known under the assumption that the MDP is  $\text{ergodic}$ . Under this assumption, we introduce IMED-RL and prove that its regret upper bound asymptotically matches the regret lower bound. We discuss both the case when the supports of transitions are unknown, and the more informative but a priori harder-to-exploit-optimal case when they are known. Rewards are assumed light-tailed, semi-bounded from above. Last, we provide numerical illustrations on classical tabular MDPs,  $\text{ergodic}$  and  $\text{communicative}$  only, showing the competitiveness of IMED-RL in finite-time against state-of-the-art algorithms. IMED-RL also benefits from a lighter complexity.

## [Asymmetric Temperature Scaling Makes Larger Networks Teach Well Again](#)

- Xin-Chun Li · Wen-shu Fan · Shaoming Song · Yinchuan Li · bingshuai Li · Shao Yunfeng · De-Chuan Zhan
- abstract@[open-review](#): Knowledge Distillation (KD) aims at transferring the knowledge of a well-performed neural network (the  $\{\text{it teacher}\}$ ) to a weaker one (the  $\{\text{it student}\}$ ). A peculiar phenomenon is that a more accurate model doesn't necessarily teach better, and temperature adjustment can neither alleviate the mismatched capacity. To explain this, we decompose the efficacy of KD into three parts:  $\{\text{it correct guidance}\}$ ,  $\{\text{it smooth regularization}\}$ , and  $\{\text{it class discriminability}\}$ . The last term describes the distinctness of  $\{\text{it wrong class probabilities}\}$  that the teacher provides in KD. Complex teachers tend to be over-confident and traditional temperature scaling limits the efficacy of  $\{\text{it class discriminability}\}$ , resulting in less discriminative wrong class probabilities. Therefore, we propose  $\{\text{it Asymmetric Temperature Scaling (ATS)}\}$ , which separately applies a higher/lower temperature to the correct/wrong class. ATS enlarges the variance of wrong class probabilities in the teacher's label and makes the students grasp the absolute affinities of wrong classes to the target class as discriminative as possible. Both theoretical analysis and extensive experimental results demonstrate the effectiveness of ATS.

## [Towards Versatile Embodied Navigation](#)

- Hanqing Wang · Wei Liang · Luc V Gool · Wenguan Wang
- abstract@[open-review](#): With the emergence of varied visual navigation tasks (e.g., image-/object-/audio-goal and vision-language navigation) that specify the target in different ways, the community has made appealing advances in training specialized agents capable of handling individual navigation tasks well. Given plenty of embodied navigation tasks and task-specific solutions, we address a more fundamental question: can we learn a single powerful agent that masters not one but multiple navigation tasks concurrently? First, we propose VXN, a large-scale 3D dataset that instantiates four classic navigation tasks in standardized, continuous, and audiovisual-rich environments. Second, we propose Vienna, a versatile embodied navigation agent that simultaneously learns to perform the four navigation tasks with one model. Building upon a full-attentive architecture, Vienna formulates various navigation tasks as a unified, parse-and-query procedure: the target description, augmented with four task embeddings, is comprehensively interpreted into a set of diversified goal vectors, which are refined as the navigation progresses, and used as queries to retrieve supportive context from episodic history for decision making. This enables the reuse of knowledge across navigation tasks with varying input domains/modalities. We empirically demonstrate that, compared with learning each visual navigation task individually, our multitask agent achieves comparable or even better performance with reduced complexity. Our dataset and code will be released.

## [Laplacian Autoencoders for Learning Stochastic Representations](#)

- Marco Miani · Frederik Warburg · Pablo Moreno-Muñoz · Nicki Skafte · Søren Hauberg
- abstract@[open-review](#): Established methods for unsupervised representation learning such as variational autoencoders produce none or poorly calibrated uncertainty estimates making it difficult to evaluate if learned representations are stable and reliable. In this work, we present a Bayesian autoencoder for unsupervised representation learning, which is trained using a novel variational lower-bound of the autoencoder evidence. This is maximized using Monte Carlo EM with a variational distribution that takes the shape of a Laplace approximation. We develop a new Hessian approximation that scales linearly with data size allowing us to model high-dimensional data. Empirically, we show that our Laplacian autoencoder estimates well-calibrated uncertainties in both latent and output space. We demonstrate that this results in improved performance across a multitude of downstream tasks.

## [Identify and Remove Backdoor Neurons through Clean-Poisoned Mixture Distribution](#)

- Runkai Zheng · Rongjun Tang · Jianze Li · Li Liu
- abstract@[open-review](#): Convolutional neural networks (CNN) can be manipulated to perform specific behavior when encountering a particular trigger pattern without affecting the performance on normal samples, which is referred to as backdoor attack. Backdoor attack is usually achieved by injecting a small proportion of poisoned samples into the training set, through which the victim trains a model embedded with the designated backdoor. In this work, we demonstrate that the backdoor neurons in an infected neural network have a mixture of two distributions with significantly different moments, formed by benign samples and poisoned samples, respectively. This property is shown to be attack-invariant and allow us to efficiently locate the backdoor neurons. On this basis, we make several realistic assumptions on the neuron activation distributions and propose two backdoor neuron detection strategies based on (1) the differential entropy of the neurons and (2) the KL divergence between the benign sample distribution and a poisoned statistics based hypothetical distribution. Experimental results show that our proposed defense strategies are both efficient and effective against various backdoor attacks.

## [Star Temporal Classification: Sequence Modeling with Partially Labeled Data](#)

- Vineel Pratap · Awni Hannun · Gabriel Synnaeve · Ronan Collobert
- abstract@[open-review](#): We develop an algorithm which can learn from partially labeled and unsegmented sequential data. Most sequential loss functions, such as Connectionist Temporal Classification (CTC), break down when many labels are missing. We address this problem with Star Temporal Classification (STC) which uses a special star token to allow alignments which include all possible tokens whenever a token could be missing. We express STC as the composition of weighted finite-state transducers (WFSTs) and use GTN (a framework for automatic differentiation with WFSTs) to compute gradients. We perform extensive experiments on automatic speech recognition. These experiments show that STC can close the performance gap with supervised baseline to about 1% WER when up to 70% of the labels are missing. We also perform experiments in handwriting recognition to show that our method easily applies to other temporal classification tasks.

## [Generalized Variational Inference in Function Spaces: Gaussian Measures meet Bayesian Deep Learning](#)

- Veit David Wild · Robert Hu · Dino Sejdinovic
- abstract@[open-review](#): We develop a framework for generalized variational inference in infinite-dimensional function spaces and use it to construct a method termed Gaussian Wasserstein inference (GWI). GWI leverages the Wasserstein distance between Gaussian measures on the Hilbert space of square-integrable functions in order to determine a variational posterior using a tractable optimization criterion and avoids pathologies arising in standard variational function space inference. An exciting application of GWI is the ability to use deep neural networks in the variational parametrization of GWI, combining their superior predictive performance with the principled uncertainty quantification analogous to that of Gaussian processes. The proposed method obtains state-of-the-art performance on several benchmark datasets.

## [Explaining Preferences with Shapley Values](#)

- Robert Hu · Siu Lun Chau · Jaime Ferrando Huertas · Dino Sejdinovic
- abstract@[open-review](#): While preference modelling is becoming one of the pillars of machine learning, the problem of preference explanation remains challenging and underexplored. In this paper, we propose \textsc{Pref-SHAP}, a Shapley value-based model explanation framework for pairwise comparison data. We derive the appropriate value functions for preference models and further extend the framework to model and explain \emph{context specific} information, such as the surface type in a tennis game. To demonstrate the utility of \textsc{Pref-SHAP}, we apply our method to a variety of synthetic and real-world datasets and show that richer and more insightful explanations can be obtained over the baseline.

## [Uncertainty-Aware Hierarchical Refinement for Incremental Implicitly-Refined Classification](#)

- Jian Yang · Kai Zhu · Kecheng Zheng · Yang Cao
- abstract@[open-review](#): Incremental implicitly-refined classification aims at assigning hierarchical labels to the same instance encountered at different tasks. Existing methods tend to fail in generating hierarchy-invariant descriptor when novel classes are inherited from the old ones. To address the issue, this paper explores the inheritance relations in the process of multi-level semantic increment, and propose an Uncertainty-Aware Hierarchical Refinement (UAHR) scheme. Our proposed scheme consists of a global representation extension strategy that enhances the discrimination of incremental representation by widening the corresponding margin distance, and a hierarchical distribution alignment strategy that refines the distillation process by explicitly determining the inheritance relationship of the incremental class. Particularly, the shifting subclasses are corrected under the guidance of hierarchical uncertainty, ensuring the consistency of the homogeneous features. Extensive experiments on benchmarks IIRC-CIFAR and IIRC-ImageNet-lite demonstrate the superiority of our proposed method over state-of-the-art.

## [SwinTrack: A Simple and Strong Baseline for Transformer Tracking](#)

- Liting Lin · Heng Fan · Zhipeng Zhang · Yong Xu · Haibin Ling
- abstract@[open-review](#): Recently Transformer has been largely explored in tracking and shown state-of-the-art (SOTA) performance. However, existing efforts mainly focus on fusing and enhancing features generated by convolutional neural networks (CNNs). The potential of Transformer in representation learning remains under-explored. In this paper, we aim to further unleash the power of Transformer by proposing a simple yet efficient fully-attentional tracker, dubbed \textbf{SwinTrack}, within classic Siamese framework. In particular, both representation learning and feature fusion in SwinTrack leverage the Transformer architecture, enabling better feature interactions for tracking than pure CNN or hybrid CNN-Transformer frameworks. Besides, to further enhance robustness, we present a novel motion token that embeds historical target trajectory to improve tracking by providing temporal context. Our motion token is lightweight with negligible computation but brings clear gains. In our thorough experiments, SwinTrack exceeds existing approaches on multiple benchmarks. Particularly, on the challenging LaSOT, SwinTrack sets a new record with \textbf{0.713} SUC score. It also achieves SOTA results on other benchmarks. We expect SwinTrack to serve as a solid baseline for Transformer tracking and facilitate future research. Our codes and results will be released.

## [Proppo: a Message Passing Framework for Customizable and Composable Learning Algorithms](#)

- Paavo Parmas · Takuma Seno
- abstract@[open-review](#): While existing automatic differentiation (AD) frameworks allow flexibly composing model architectures, they do not provide the same flexibility for composing learning algorithms---everything has to be implemented in terms of back propagation. To address this gap, we invent Automatic Propagation (AP) software, which generalizes AD, and allows custom and composable construction of complex learning algorithms. The framework allows packaging custom learning algorithms into propagators that automatically implement the necessary computations, and can be reused across different computation graphs. We implement Proppo, a prototype AP software package built on top of the Pytorch AD framework. To demonstrate the utility of Proppo, we use it to implement Monte Carlo gradient estimation techniques, such as reparameterization and likelihood ratio gradients, as well as the total propagation algorithm and Gaussian shaping gradients, which were previously used in model-based reinforcement learning, but do not have any publicly available implementation. Finally, in minimalistic experiments, we show that these methods allow increasing the gradient accuracy by orders of magnitude, particularly when the machine learning system is at the edge of chaos.

## [C2FAR: Coarse-to-Fine Autoregressive Networks for Precise Probabilistic Forecasting](#)

- Shane Bergsma · Tim Zeyl · Javad Rahimipour Anaraki · Lei Guo
- abstract@[open-review](#): We present coarse-to-fine autoregressive networks (C2FAR), a method for modeling the probability distribution of univariate random variables. In C2FAR, a coarse-to-fine discretization of the variable is generated autoregressively; progressively finer intervals of support are generated, conditioned on previously-generated coarser intervals. Unlike prior binned distributions, C2FAR can represent values with exponentially higher precision, for only a linear increase in complexity. We use C2FAR for probabilistic forecasting via a recurrent neural network, thus modeling time series autoregressively in both space and time. C2FAR is the first method to simultaneously handle discrete and continuous series of arbitrary scale and distribution shape. This flexibility enables a variety of time series use cases, including anomaly detection, interpolation, and compression. C2FAR achieves improvements over the state-of-the-art on several benchmark forecasting datasets.

## [Pitfalls of Epistemic Uncertainty Quantification through Loss Minimisation](#)

- Viktor Bengs · Eyke Höllemeier · Willem Waegeman
- abstract@[open-review](#): Uncertainty quantification has received increasing attention in machine learning in the recent past. In particular, a distinction between aleatoric and epistemic uncertainty has been found useful in this regard. The latter refers to the learner's (lack of) knowledge and appears to be especially difficult to measure and quantify. In this paper, we analyse a recent proposal based on the idea of a second-order learner, which yields predictions in the form of distributions over probability distributions. While standard (first-order) learners can be trained to predict accurate probabilities, namely by minimising suitable loss functions on sample data, we show that loss minimisation does not work for second-order predictors: The loss functions proposed for inducing such predictors do not incentivise the learner to represent its epistemic uncertainty in a faithful way.

## [\\$\\textit{Public Wisdom Matters!}\\\\$ Discourse-Aware Hyperbolic Fourier Co-Attention for Social Text Classification](#)

- Karish Grover · S M Phaneendra Angara · Md Shad Akhtar · Tanmoy Chakraborty
- abstract@[open-review](#): Social media has become the fulcrum of all forms of communication. Classifying social texts such as fake news, rumour, sarcasm, etc. has gained significant attention. The surface-level signals expressed by a social-text itself may not be adequate for such tasks; therefore, recent methods attempted to incorporate other intrinsic signals such as user behavior and the underlying graph structure. Oftentimes, the 'public wisdom' expressed through the comments/replies to a social-text acts as a surrogate of crowd-sourced view and may provide us with complementary signals. State-of-the-art methods on social-text classification tend to ignore such a rich hierarchical signal. Here, we propose \\$\textbf{\\texttt{Hyphen}}\\$, a discourse-aware hyperbolic spectral co-attention network. \\$\textbf{\\texttt{Hyphen}}\\$ is a fusion of hyperbolic graph representation learning with a novel Fourier co-attention mechanism in an attempt to \\$\\textit{generalise}\\$ the social-text classification tasks by incorporating \\$\\textit{public discourse}\\$. We parse public discourse as an Abstract Meaning Representation (AMR) graph and use the powerful hyperbolic geometric representation to model graphs with hierarchical structure. Finally, we equip it with a novel Fourier co-attention mechanism to capture the correlation between the source post and public discourse. Extensive experiments on four different social-text classification tasks, namely detecting fake news, hate speech, rumour, and sarcasm, show that \\$\textbf{\\texttt{Hyphen}}\\$ generalises well, and achieves state-of-the-art results on ten benchmark datasets. We also employ a sentence-level fact-checked and annotated dataset to evaluate how \\$\textbf{\\texttt{Hyphen}}\\$ is capable of producing \\$\\textit{explanations}\\$ as analogous evidence to the final prediction.

## [Exploiting Reward Shifting in Value-Based Deep RL](#)

- Hao Sun · Lei Han · Rui Yang · Xiaoteng Ma · Jian Guo · Bolei Zhou
- abstract@[open-review](#): In this work, we study the simple yet universally applicable case of reward shaping in value-based Deep Reinforcement Learning (DRL). We show that reward shifting in the form of the linear transformation is equivalent to changing the initialization of the \\$Q\\$-function in function approximation. Based on such an equivalence, we bring the key insight that a positive reward shifting leads to conservative exploitation, while a negative reward shifting leads to curiosity-driven exploration. Accordingly, conservative exploitation improves offline RL value estimation, and optimistic value estimation improves exploration for online RL. We validate our insight on a range of RL tasks and show its improvement over baselines: (1) In offline RL, the conservative exploitation leads to improved performance based on off-the-shelf algorithms; (2) In online continuous control, multiple value functions with different shifting constants can be used to tackle the exploration-exploitation dilemma for better sample efficiency; (3) In discrete control tasks, a negative reward shifting yields an improvement over the curiosity-based exploration method.

## [Emergent Communication: Generalization and Overfitting in Lewis Games](#)

- Mathieu Rita · Corentin Tallec · Paul Michel · Jean-Bastien Grill · Olivier Pietquin · Emmanuel Dupoux · Florian Strub
- abstract@[open-review](#): Lewis signaling games are a class of simple communication games for simulating the emergence of language. In these games, two agents must agree on a communication protocol in order to solve a cooperative task. Previous work has shown that agents trained to play this game with reinforcement learning tend to develop languages that display undesirable properties from a linguistic point of view (lack of generalization, lack of compositionality, etc). In this paper, we aim to provide better understanding of this phenomenon by analytically studying the learning problem in Lewis

games. As a core contribution, we demonstrate that the standard objective in Lewis games can be decomposed in two components: a co-adaptation loss and an information loss. This decomposition enables us to surface two potential sources of overfitting, which we show may undermine the emergence of a structured communication protocol. In particular, when we control for overfitting on the co-adaptation loss, we recover desired properties in the emergent languages: they are more compositional and generalize better.

## [Fast Algorithms for Packing Proportional Fairness and its Dual](#)

- Francisco Criado · David Martínez-Rubio · Sebastian Pokutta
- abstract@[open-review](#): The proportional fair resource allocation problem is a major problem studied in flow control of networks, operations research, and economic theory, where it has found numerous applications. This problem, defined as the constrained maximization of  $\sum_i \log x_i$ , is known as the packing proportional fairness problem when the feasible set is defined by positive linear constraints and  $x \in \mathbb{R}^n$ . In this work, we present a distributed accelerated first-order method for this problem which improves upon previous approaches. We also design an algorithm for the optimization of its dual problem. Both algorithms are width-independent. Finally, we show the latter problem has applications to the volume reduction of bounding simplices in an old linear programming algorithm of [yamnitsky1982](#), and we obtain some improvements as a result.

## [Is this the Right Neighborhood? Accurate and Query Efficient Model Agnostic Explanations](#)

- Amit Dhurandhar · Karthikeyan Natesan Ramamurthy · Karthikeyan Shanmugam
- abstract@[open-review](#): There have been multiple works that try to ascertain explanations for decisions of black box models on particular inputs by perturbing the input or by sampling around it, creating a neighborhood and then fitting a sparse (linear) model (e.g. LIME). Many of these methods are unstable and so more recent work tries to find stable or robust alternatives. However, stable solutions may not accurately represent the behavior of the model around the input. Thus, the question we ask in this paper is are we approximating the local boundary around the input accurately? In particular, are we sampling the right neighborhood so that a linear approximation of the black box is faithful to its true behavior around that input given that the black box can be highly non-linear (viz. deep relu network with many linear pieces). It is difficult to know the correct neighborhood width (or radius) as too small a width can lead to a bad condition number of the inverse covariance matrix of function fitting procedures resulting in unstable predictions, while too large a width may lead to accounting for multiple linear pieces and consequently a poor local approximation. We in this paper propose a simple approach that is robust across neighborhood widths in recovering faithful local explanations. In addition to a naive implementation of our approach which can still be accurate, we propose a novel adaptive neighborhood sampling scheme (ANS) that we formally show can be much more sample and query efficient. We then empirically evaluate our approach on real data where our explanations are significantly more sample and query efficient than the competitors, while also being faithful and stable across different widths.

## [Latency-aware Spatial-wise Dynamic Networks](#)

- Yizeng Han · Zhihang Yuan · Yifan Pu · Chenhao Xue · Shiji Song · Guangyu Sun · Gao Huang
- abstract@[open-review](#): Spatial-wise dynamic convolution has become a promising approach to improving the inference efficiency of deep networks. By allocating more computation to the most informative feature pixels, such an adaptive inference paradigm alleviates the spatial redundancy in image features and reduces a considerable amount of unnecessary computation. However, the theoretical efficiency achieved by previous methods can hardly translate into the realistic speedup, especially on the multi-core processors (e.g. GPUs). The key challenge is that the existing literature has only focused on designing algorithms with minimal computation, ignoring the fact that the practical latency can also be influenced by scheduling strategies and hardware properties. To bridge the gap between the theoretical computation and the practical efficiency, we propose a latency-aware spatial-wise dynamic network (LASNet), which performs *coarse-grained* spatially adaptive inference under the guidance of a novel latency prediction model. This latency prediction model can efficiently estimate the inference latency of dynamic networks by simultaneously considering the algorithms, the scheduling strategies, and the hardware properties. We use the latency predictor to guide both the algorithm design and the scheduling optimization on various hardware platforms. Experiments on image classification demonstrate that the proposed framework significantly improves the trade-off between the accuracy and the inference efficiency of deep networks. For example, the average latency of a ResNet-101 on the ImageNet validation set could be reduced by 23% and 45% on a server GPU (Nvidia Tesla-V100) and an IoT device (Nvidia Jetson TX2 GPU) respectively without sacrificing the accuracy.

## [GALOIS: Boosting Deep Reinforcement Learning via Generalizable Logic Synthesis](#)

- Yushi Cao · Zhiming Li · Tianpei Yang · Hao Zhang · YAN ZHENG · Yi Li · Jianye Hao · Yang Liu
- abstract@[open-review](#): Despite achieving superior performance in human-level control problems, unlike humans, deep reinforcement learning (DRL) lacks high-order intelligence (e.g., logic deduction and reuse), thus it behaves ineffectively than humans regarding learning and generalization in complex problems. Previous works attempt to directly synthesize a white-box logic program as the DRL policy, manifesting logic-driven behaviors. However, most synthesis methods are built on imperative or declarative programming, and each has a distinct limitation, respectively. The former ignores the cause-effect logic during synthesis, resulting in low generalizability across tasks. The latter is strictly proof-based, thus failing to synthesize programs with complex hierarchical logic. In this paper, we combine the above two paradigms together and propose a novel Generalizable Logic Synthesis (GALOIS) framework to synthesize hierarchical and strict cause-effect logic programs. GALOIS leverages the program sketch and defines a new sketch-based hybrid program language for guiding the synthesis. Based on that, GALOIS proposes a sketch-based program synthesis method to automatically generate white-box programs with generalizable and interpretable cause-effect logic. Extensive evaluations on various decision-making tasks with complex logic demonstrate the superiority of GALOIS over mainstream baselines regarding the asymptotic performance, generalizability, and great knowledge reusability across different environments.

## [What is a Good Metric to Study Generalization of Minimax Learners?](#)

- Asuman Ozdaglar · Sarah Pattathil · Jiawei Zhang · Kaiqing Zhang
- abstract@[open-review](#): Minimax optimization has served as the backbone of many machine learning (ML) problems. Although the convergence behavior of optimization algorithms has been extensively studied in minimax settings, their generalization guarantees, i.e., how the model trained on empirical data performs on the unseen testing data, have been relatively under-explored. A fundamental question remains elusive: What is a good metric to study generalization of minimax learners? In this paper, we aim to answer this question by first showing that primal risk, a universal metric to study generalization in minimization problems, fails in simple examples of minimax problems. Furthermore, another popular metric, the primal-dual risk, also fails to characterize the generalization behavior for minimax problems with nonconvexity, due to non-existence of saddle points. We thus propose a new metric to study generalization of minimax learners: the primal gap, to circumvent these issues. Next, we derive generalization bounds for the primal gap in nonconvex-concave settings. As byproducts of our analysis, we also solve two open questions: establishing generalization bounds for primal risk and primal-dual risk in this setting, and in the strong sense, i.e., without assuming that the maximization and expectation can be interchanged. Finally, we leverage this new metric to compare the generalization behavior of two popular algorithms - gradient descent-ascent (GDA) and gradient descent-max (GDMax) in minimax optimization.

## [Accelerated Primal-Dual Gradient Method for Smooth and Convex-Concave Saddle-Point Problems with Bilinear Coupling](#)

- Dmitry Kovalev · Alexander Gasnikov · Peter Richtarik

- abstract@[open-review](#): In this paper we study the convex-concave saddle-point problem  $\min_x \max_y f(x) + y^\top \mathbf{A}x - g(y)$ , where  $f(x)$  and  $g(y)$  are smooth and convex functions. We propose an Accelerated Primal-Dual Gradient Method (APDG) for solving this problem, achieving (i) an optimal linear convergence rate in the strongly-convex-strongly-concave regime, matching the lower complexity bound (Zhang et al., 2021), and (ii) an accelerated linear convergence rate in the case when only one of the functions  $f(x)$  and  $g(y)$  is strongly convex or even none of them are. Finally, we obtain a linearly convergent algorithm for the general smooth and convex-concave saddle point problem  $\min_x \max_y F(x,y)$  without the requirement of strong convexity or strong concavity.

## [Communication Acceleration of Local Gradient Methods via an Accelerated Primal-Dual Algorithm with an Inexact Prox](#)

- Abdurakhmon Sadiev · Dmitry Kovalev · Peter Richtarik
- abstract@[open-review](#): Inspired by a recent breakthrough of Mishchenko et al. [2022], who for the first time showed that local gradient steps can lead to provable communication acceleration, we propose an alternative algorithm which obtains the same communication acceleration as their method (ProxSkip). Our approach is very different, however: it is based on the celebrated method of Chambolle and Pock [2011], with several nontrivial modifications: i) we allow for an inexact computation of the prox operator of a certain smooth strongly convex function via a suitable gradient-based method (e.g., GD or Fast GD), ii) we perform a careful modification of the dual update step in order to retain linear convergence. Our general results offer the new state-of-the-art rates for the class of strongly convex-concave saddle-point problems with bilinear coupling characterized by the absence of smoothness in the dual function. When applied to federated learning, we obtain a theoretically better alternative to ProxSkip: our method requires fewer local steps ( $\mathcal{O}(\kappa^{1/3})$  or  $\mathcal{O}(\kappa^{1/4})$ , compared to  $\mathcal{O}(\kappa^{1/2})$  of ProxSkip), and performs a deterministic number of local steps instead. Like ProxSkip, our method can be applied to optimization over a connected network, and we obtain theoretical improvements here as well.

## [Reincarnating Reinforcement Learning: Reusing Prior Computation to Accelerate Progress](#)

- Rishabh Agarwal · Max Schwarzer · Pablo Samuel Castro · Aaron Courville · Marc Bellemare
- abstract@[open-review](#): Learning tabula rasa, that is without any prior knowledge, is the prevalent workflow in reinforcement learning (RL) research. However, RL systems, when applied to large-scale settings, rarely operate tabula rasa. Such large-scale systems undergo multiple design or algorithmic changes during their development cycle and use ad hoc approaches for incorporating these changes without re-training from scratch, which would have been prohibitively expensive. Additionally, the inefficiency of deep RL typically excludes researchers without access to industrial-scale resources from tackling computationally-demanding problems. To address these issues, we present reincarnating RL as an alternative workflow or class of problem settings, where prior computational work (e.g., learned policies) is reused or transferred between design iterations of an RL agent, or from one RL agent to another. As a step towards enabling reincarnating RL from any agent to any other agent, we focus on the specific setting of efficiently transferring an existing sub-optimal policy to a standalone value-based RL agent. We find that existing approaches fail in this setting and propose a simple algorithm to address their limitations. Equipped with this algorithm, we demonstrate reincarnating RL's gains over tabula rasa RL on Atari 2600 games, a challenging locomotion task, and the real-world problem of navigating stratospheric balloons. Overall, this work argues for an alternative approach to RL research, which we believe could significantly improve real-world RL adoption and help democratize it further. Open-sourced code and trained agents at [https://agarwl.github.io/reincarnating\\_rl](https://agarwl.github.io/reincarnating_rl).

## [Optimal Algorithms for Decentralized Stochastic Variational Inequalities](#)

- Dmitry Kovalev · Aleksandr Beznosikov · Abdurakhmon Sadiev · Michael Persianov · Peter Richtarik · Alexander Gasnikov
- abstract@[open-review](#): Variational inequalities are a formalism that includes games, minimization, saddle point, and equilibrium problems as special cases. Methods for variational inequalities are therefore universal approaches for many applied tasks, including machine learning problems. This work concentrates on the decentralized setting, which is increasingly important but not well understood. In particular, we consider decentralized stochastic (sum-type) variational inequalities over fixed and time-varying networks. We present lower complexity bounds for both communication and local iterations and construct optimal algorithms that match these lower bounds. Our algorithms are the best among the available literature not only in the decentralized stochastic case, but also in the decentralized deterministic and non-distributed stochastic cases. Experimental results confirm the effectiveness of the presented algorithms.

## [Logical Credal Networks](#)

- Radu Marinescu · Haifeng Qian · Alexander Gray · Debarun Bhattacharya · Francisco Barahona · Tian Gao · Ryan Riegel · Pravinda Sahu
- abstract@[open-review](#): We introduce Logical Credal Networks (or LCNs for short) -- an expressive probabilistic logic that generalizes prior formalisms that combine logic and probability. Given imprecise information represented by probability bounds and conditional probability bounds on logic formulas, an LCN specifies a set of probability distributions over all its interpretations. Our approach allows propositional and first-order logic formulas with few restrictions, e.g., without requiring acyclicity. We also define a generalized Markov condition that allows us to identify implicit independence relations between atomic formulas. We evaluate our method on benchmark problems such as random networks, Mastermind games with uncertainty and credit card fraud detection. Our results show that the LCN outperforms existing approaches; its advantage lies in aggregating multiple sources of imprecise information.

## [Knowledge-Aware Bayesian Deep Topic Model](#)

- Dongsheng Wang · Yi.shi Xu · Miaoge Li · Zhibin Duan · Chaojie Wang · Bo Chen · Mingyuan Zhou
- abstract@[open-review](#): We propose a Bayesian generative model for incorporating prior domain knowledge into hierarchical topic modeling. Although embedded topic models (ETMs) and its variants have gained promising performance in text analysis, they mainly focus on mining word co-occurrence patterns, ignoring potentially easy-to-obtain prior topic hierarchies that could help enhance topic coherence. While several knowledge-based topic models have recently been proposed, they are either only applicable to shallow hierarchies or sensitive to the quality of the provided prior knowledge. To this end, we develop a novel deep ETM that jointly models the documents and the given prior knowledge by embedding the words and topics into the same space. Guided by the provided knowledge, the proposed model tends to discover topic hierarchies that are organized into interpretable taxonomies. Besides, with a technique for adapting a given graph, our extended version allows the provided prior topic structure to be finetuned to match the target corpus. Extensive experiments show that our proposed model efficiently integrates the prior knowledge and improves both hierarchical topic discovery and document representation.

## [Stars: Tera-Scale Graph Building for Clustering and Learning](#)

- CJ Carey · Jonathan Halcrow · Rajesh Jayaram · Vahab Mirrokni · Warren Schudy · Peilin Zhong
- abstract@[open-review](#): A fundamental procedure in the analysis of massive datasets is the construction of similarity graphs. Such graphs play a key role for many downstream tasks, including clustering, classification, graph learning, and nearest neighbor search. For these tasks, it is critical to build graphs which are sparse yet still representative of the underlying data. The benefits of sparsity are twofold: firstly, constructing dense graphs is infeasible in practice for large datasets, and secondly, the runtime of downstream tasks is directly influenced by the sparsity of the similarity graph. In this work, we present **Stars**: a highly scalable method for building extremely sparse graphs via two-hop spanners, which are graphs where similar points are connected by a path of length at most two. Stars can construct two-hop spanners with significantly fewer similarity comparisons, which are a major bottleneck for learning based models where comparisons are expensive to evaluate. Theoretically, we demonstrate that Stars builds a graph in nearly-linear

time, where approximate nearest neighbors are contained within two-hop neighborhoods. In practice, we have deployed Stars for multiple data sets allowing for graph building at the \text{Tera-Scale}, i.e., for graphs with hundreds of billions of nodes and tens of trillions of edges. We evaluate the performance of Stars for clustering and graph learning, and demonstrate 10\text{times} to 1000-fold improvements in pairwise similarity comparisons and significant running time speedups with negligible quality loss.

## [Association Graph Learning for Multi-Task Classification with Category Shifts](#)

- Jiayi Shen · Zehao Xiao · Xiantong Zhen · Cees Snoek · Marcel Worring
- abstract@[open-review](#): In this paper, we focus on multi-task classification, where related classification tasks share the same label space and are learned simultaneously. In particular, we tackle a new setting, which is more realistic than currently addressed in the literature, where categories shift from training to test data. Hence, individual tasks do not contain complete training data for the categories in the test set. To generalize to such test data, it is crucial for individual tasks to leverage knowledge from related tasks. To this end, we propose learning an association graph to transfer knowledge among tasks for missing classes. We construct the association graph with nodes representing tasks, classes and instances, and encode the relationships among the nodes in the edges to guide the knowledge transfer between them. By message passing on the association graph, our model enhances the categorical information of each instance, making it more discriminative. To avoid spurious correlations between task and class nodes in the graph, we introduce an assignment entropy maximization that encourages each class node to balance its edge weights. This enables all tasks to fully utilize the categorical information from related tasks. An extensive evaluation on three general benchmarks and a medical dataset for skin lesion classification reveals that our method consistently performs better than representative baselines.

## [CoNT: Contrastive Neural Text Generation](#)

- Chenxin An · Jiangtao Feng · Kai Lv · Lingpeng Kong · Xipeng Qiu · Xuanjing Huang
- abstract@[open-review](#): Recently, contrastive learning attracts increasing interests in neural text generation as a new solution to alleviate the exposure bias problem. It introduces a sequence-level training signal which is crucial to generation tasks that always rely on auto-regressive decoding. However, previous methods using contrastive learning in neural text generation usually lead to inferior performance. In this paper, we analyse the underlying reasons and propose a new Contrastive Neural Text generation framework, CoNT. CoNT addresses bottlenecks that prevent contrastive learning from being widely adopted in generation tasks from three aspects -- the construction of contrastive examples, the choice of the contrastive loss, and the strategy in decoding. We validate CoNT on five generation tasks with ten benchmarks, including machine translation, summarization, code comment generation, data-to-text generation and commonsense generation. Experimental results show that CoNT clearly outperforms its baseline on all the ten benchmarks with a convincing margin. Especially, CoNT surpasses previous the most competitive contrastive learning method for text generation, by 1.50 BLEU on machine translation and 1.77 ROUGE-1 on summarization, respectively. It achieves new state-of-the-art on summarization, code comment generation (without external data) and data-to-text generation.

## [Revisiting Active Sets for Gaussian Process Decoders](#)

- Pablo Moreno-Muñoz · Cilie Feldager · Søren Hauberg
- abstract@[open-review](#): Decoders built on Gaussian processes (GPs) are enticing due to the marginalisation over the non-linear function space. Such models (also known as GP-LVMs) are often expensive and notoriously difficult to train in practice, but can be scaled using variational inference and inducing points. In this paper, we revisit active set approximations. We develop a new stochastic estimate of the log-marginal likelihood based on recently discovered links to cross-validation, and propose a computationally efficient approximation thereof. We demonstrate that the resulting stochastic active sets (SAS) approximation significantly improves the robustness of GP decoder training while reducing computational cost. The SAS-GP obtains more structure in the latent space, scales to many datapoints and learns better representations than variational autoencoders, which is rarely the case for GP decoders.

## [Contrastive Language-Image Pre-Training with Knowledge Graphs](#)

- Xuran Pan · Tianzhu Ye · Dongchen Han · Shiji Song · Gao Huang
- abstract@[open-review](#): Recent years have witnessed the vast development of large-scale pre-training frameworks that can extract multi-modal representations in a unified form and achieve promising performances when transferred to downstream tasks. Nevertheless, existing approaches mainly focus on pre-training with simple image-text pairs, while neglecting the semantic connections between concepts from different modalities. In this paper, we propose a knowledge-based pre-training framework, dubbed \text{Knowledge-CLIP}, that injects semantic information into the widely used CLIP model. Through introducing knowledge-based objectives in the pre-training process and utilizing different types of knowledge graphs as training data, our model can semantically align the representations in vision and language, and also enhance the reasoning ability across scenarios and modalities. Extensive experiments on various vision-language downstream tasks demonstrate the effectiveness of Knowledge-CLIP comparing with the original CLIP and competitive baselines.

## [Isolating and Leveraging Controllable and Noncontrollable Visual Dynamics in World Models](#)

- Minting Pan · Xiangming Zhu · Yunbo Wang · Xiaokang Yang
- abstract@[open-review](#): World models learn the consequences of actions in vision-based interactive systems. However, in practical scenarios such as autonomous driving, there commonly exists noncontrollable dynamics independent of the action signals, making it difficult to learn effective world models. Naturally, therefore, we need to enable the world models to decouple the controllable and noncontrollable dynamics from the entangled spatiotemporal data. To this end, we present a reinforcement learning approach named Iso-Dream, which expands the Dream-to-Control framework in two aspects. First, the world model contains a three-branch neural architecture. By solving the inverse dynamics problem, it learns to factorize latent representations according to the responses to action signals. Second, in the process of behavior learning, we estimate the state values by rolling-out a sequence of noncontrollable states (less related to the actions) into the future and associate the current controllable state with them. In this way, the isolation of mixed dynamics can greatly facilitate long-horizon decision-making tasks in realistic scenes, such as avoiding potential future risks by predicting the movement of other vehicles in autonomous driving. Experiments show that Iso-Dream is effective in decoupling the mixed dynamics and remarkably outperforms existing approaches in a wide range of visual control and prediction domains.

## [Self-Supervised Contrastive Pre-Training For Time Series via Time-Frequency Consistency](#)

- Xiang Zhang · Ziyuan Zhao · Theodoros Tsiligkaridis · Marinka Zitnik
- abstract@[open-review](#): Time series datasets pose a unique challenge for pre-training due to various kinds of discrepancy between pre-training and target domains, such as shifts in temporal dynamics, fast-evolving trends, long-range and short cyclic effects. While domain adaptation methods can mitigate these shifts, most methods need examples directly from the target domain, making them suboptimal for pre-training on time series. To fill this gap, methods need to accommodate a set of target domains with different temporal dynamics and be capable to do so without seeing any target examples during pre-training. Relative to other fields, in time series we expect that time-based and frequency-based representations of the same example and their local augmentations are located close together in the time-frequency space. To this end, we posit that time-frequency consistency (TF-C) --- embedding a time-based neighborhood of a particular example close to its frequency-based neighborhood and back --- is desirable for pre-training. Motivated by TF-C, we optimize a decomposable pre-training model, where the self-supervised signal is provided by the distance between time and frequency components, each individually trained by contrastive estimation. We evaluate the new method on eight datasets including electrodiagnostic testing, human daily activity

recognition, mechanical fault detection, and physical status monitoring. Experiments against eight state-of-the-art methods show our method outperforms the baselines by 15.4% (F1 score) on average in one-to-one settings (e.g., fine-tuning an EEG-pretrained model on EMG data) and by up to 8.4% (F1 score) in challenging one-to-many settings (e.g., fine-tuning an EEG-pretrained model for either hand-gesture recognition or mechanical fault prediction), reflecting the breadth of scenarios that arise in real-world applications. The source code and all datasets are available at <https://anonymous.4open.science/r/TFC-pretraining-6B07>.

## [DTG-SSOD: Dense Teacher Guidance for Semi-Supervised Object Detection](#)

- Gang Li · Xiang Li · Yujie Wang · Shanshan Zhang · Wu Yichao · Ding Liang
- abstract@[open-review](#): The Mean-Teacher (MT) scheme is widely adopted in semi-supervised object detection (SSOD). In MT, sparse pseudo labels, offered by the final predictions of the teacher (e.g., after Non Maximum Suppression (NMS) post-processing), are adopted for the dense supervision for the student via hand-crafted label assignment. However, the "sparse-to-dense" paradigm complicates the pipeline of SSOD, and simultaneously neglects the powerful direct, dense teacher supervision. In this paper, we attempt to directly leverage the dense guidance of teacher to supervise student training, i.e., the "dense-to-dense" paradigm. Specifically, we propose the Inverse NMS Clustering (INC) and Rank Matching (RM) to instantiate the dense supervision, without the widely used, conventional sparse pseudo labels. INC leads the student to group candidate boxes into clusters in NMS as the teacher does, which is implemented by learning grouping information revealed in NMS procedure of the teacher. After obtaining the same grouping scheme as the teacher via INC, the student further imitates the rank distribution of the teacher over clustered candidates through Rank Matching. With the proposed INC and RM, we integrate Dense Teacher Guidance into Semi-Supervised Object Detection (termed "DTG-SSOD"), successfully abandoning sparse pseudo labels and enabling more informative learning on unlabeled data. On COCO benchmark, our DTG-SSOD achieves state-of-the-art performance under various labelling ratios. For example, under 10% labelling ratio, DTG-SSOD improves the supervised baseline from 26.9 to 35.9 mAP, outperforming the previous best method Soft Teacher by 1.9 points.

## [Model-Based Opponent Modeling](#)

- XiaoPeng Yu · Jiechuan Jiang · Wanpeng Zhang · Haobin Jiang · Zongqing Lu
- abstract@[open-review](#): When one agent interacts with a multi-agent environment, it is challenging to deal with various opponents unseen before. Modeling the behaviors, goals, or beliefs of opponents could help the agent adjust its policy to adapt to different opponents. In addition, it is also important to consider opponents who are learning simultaneously or capable of reasoning. However, existing work usually tackles only one of the aforementioned types of opponents. In this paper, we propose model-based opponent modeling (MBOM), which employs the environment model to adapt to all kinds of opponents. MBOM simulates the recursive reasoning process in the environment model and imagines a set of improving opponent policies. To effectively and accurately represent the opponent policy, MBOM further mixes the imagined opponent policies according to the similarity with the real behaviors of opponents. Empirically, we show that MBOM achieves more effective adaptation than existing methods in a variety of tasks, respectively with different types of opponents, i.e., fixed policy, naive learner, and reasoning learner.

## [Generative Time Series Forecasting with Diffusion, Denoise and Disentanglement](#)

- Yan Li · Xinjiang Lu · Yaqing Wang · Dejing Dou
- abstract@[open-review](#): Time series forecasting has been a widely explored task that is of great importance in many applications. However, it is common that real-world time series data are recorded in a short time period, which results in a big gap between the deep model and the limited and noisy time series. In this work, we propose to address the time series forecasting problem with generative modeling. and propose a bidirectional variational auto-encoder (BVAE) equipped with diffusion, denoise, and disentanglement, namely D3VAE. Specifically, a coupled diffusion probabilistic model is proposed to augment the time series data without increasing the aleatoric uncertainty contributed to the data. To ensure the generated series move towards the true target, we further propose to adapt and integrate the multiscale denoising score matching into the diffusion process for time series forecasting. In addition, to enhance the interpretability and stability of the prediction, we treat the latent variable in a multivariate manner and disentangle them on top of minimizing total correlation. Extensive experiments on both synthetic data and real-world data show that D3VAE outperforms competitive algorithms with remarkable margins.

## [Graphein - a Python Library for Geometric Deep Learning and Network Analysis on Biomolecular Structures and Interaction Networks](#)

- Arian Jamasb · Ramon Viñals Torné · Eric Ma · Yuanqi Du · Charles Harris · Kexin Huang · Dominic Hall · Pietro Liñan · Tom Blundell
- abstract@[open-review](#): Geometric deep learning has broad applications in biology, a domain where relational structure in data is often intrinsic to modelling the underlying phenomena. Currently, efforts in both geometric deep learning and, more broadly, deep learning applied to biomolecular tasks have been hampered by a scarcity of appropriate datasets accessible to domain specialists and machine learning researchers alike. To address this, we introduce Graphein as a turn-key tool for transforming raw data from widely-used bioinformatics databases into machine learning-ready datasets in a high-throughput and flexible manner. Graphein is a Python library for constructing graph and surface-mesh representations of biomolecular structures, such as proteins, nucleic acids and small molecules, and biological interaction networks for computational analysis and machine learning. Graphein provides utilities for data retrieval from widely-used bioinformatics databases for structural data, including the Protein Data Bank, the AlphaFold Structure Database, chemical data from ZINC and ChEMBL, and for biomolecular interaction networks from STRINGdb, BioGrid, TRRUST and RegNetwork. The library interfaces with popular geometric deep learning libraries: DGL, Jraph, PyTorch Geometric and PyTorch3D though remains framework agnostic as it is built on top of the PyData ecosystem to enable inter-operability with scientific computing tools and libraries. Graphein is designed to be highly flexible, allowing the user to specify each step of the data preparation, scalable to facilitate working with large protein complexes and interaction graphs, and contains useful pre-processing tools for preparing experimental files. Graphein facilitates network-based, graph-theoretic and topological analyses of structural and interaction datasets in a high-throughput manner. We envision that Graphein will facilitate developments in computational biology, graph representation learning and drug discovery. Availability and implementation: Graphein is written in Python. Source code, example usage and tutorials, datasets, and documentation are made freely available under the MIT License at the following URL: <https://anonymous.4open.science/r/graphein-3472/README.md>

## [ViTPose: Simple Vision Transformer Baselines for Human Pose Estimation](#)

- Yufei Xu · Jing Zhang · Qiming ZHANG · Dacheng Tao
- abstract@[open-review](#): Although no specific domain knowledge is considered in the design, plain vision transformers have shown excellent performance in visual recognition tasks. However, little effort has been made to reveal the potential of such simple structures for pose estimation tasks. In this paper, we show the surprisingly good capabilities of plain vision transformers for pose estimation from various aspects, namely simplicity in model structure, scalability in model size, flexibility in training paradigm, and transferability of knowledge between models, through a simple baseline model called ViTPose. Specifically, ViTPose employs plain and non-hierarchical vision transformers as backbones to extract features for a given person instance and a lightweight decoder for pose estimation. It can be scaled up from 100M to 1B parameters by taking the advantages of the scalable model capacity and high parallelism of transformers, setting a new Pareto front between throughput and performance. Besides, ViTPose is very flexible regarding the attention type, input resolution, pre-training and finetuning strategy, as well as dealing with multiple pose tasks. We also empirically demonstrate that the knowledge of large ViTPose models can be easily transferred to small ones via a simple knowledge token. Experimental results show that our basic ViTPose model outperforms representative methods on the challenging MS COCO Keypoint Detection benchmark, while the largest model sets a new state-of-the-art. The code and models will be released.

## [Entropy-Driven Mixed-Precision Quantization for Deep Network Design](#)

- Zhenhong Sun · Ce Ge · Junyan Wang · Ming Lin · Hesen Chen · Hao Li · Xiuyu Sun
- abstract@[open-review](#): Deploying deep convolutional neural networks on Internet-of-Things (IoT) devices is challenging due to the limited computational resources, such as limited SRAM memory and Flash storage. Previous works re-design a small network for IoT devices, and then compress the network size by mixed-precision quantization. This two-stage procedure cannot optimize the architecture and the corresponding quantization jointly, leading to sub-optimal tiny deep models. In this work, we propose a one-stage solution that optimizes both jointly and automatically. The key idea of our approach is to cast the joint architecture design and quantization as an Entropy Maximization process. Particularly, our algorithm automatically designs a tiny deep model such that: 1) Its representation capacity measured by entropy is maximized under the given computational budget; 2) Each layer is assigned with a proper quantization precision; 3) The overall design loop can be done on CPU, and no GPU is required. More impressively, our method can directly search high-expressiveness architecture for IoT devices within less than half a CPU hour. Extensive experiments on three widely adopted benchmarks, ImageNet, VW and WIDER FACE, demonstrate that our method can achieve the state-of-the-art performance in the tiny deep model regime. Code and pre-trained models are available at <https://github.com/alibaba/lightweight-neural-architecture-search>.

## [Towards Learning Universal Hyperparameter Optimizers with Transformers](#)

- Yutian Chen · Xingyou Song · Chansoo Lee · Zi Wang · Richard Zhang · David Dohan · Kazuya Kawakami · Greg Kochanski · Arnaud Doucet · Marc'Aurelio Ranzato · Sagi Perel · Nando de Freitas
- abstract@[open-review](#): Meta-learning hyperparameter optimization (HPO) algorithms from prior experiments is a promising approach to improve optimization efficiency over objective functions from a similar distribution. However, existing methods are restricted to learning from experiments sharing the same set of hyperparameters. In this paper, we introduce the OptFormer, the first text-based Transformer HPO framework that provides a universal end-to-end interface for jointly learning policy and function prediction when trained on vast tuning data from the wild, such as Google's Vizier database, one of the world's largest HPO datasets. Our extensive experiments demonstrate that the OptFormer can simultaneously imitate at least 7 different HPO algorithms, which can be further improved via its function uncertainty estimates. Compared to a Gaussian Process, the OptFormer also learns a robust prior distribution for hyperparameter response functions, and can thereby provide more accurate and better calibrated predictions. This work paves the path to future extensions for training a Transformer-based model as a general HPO optimizer.

## [Is Out-of-distribution Detection Learnable?](#)

- Zhen Fang · Yixuan Li · Jie Lu · Jiahua Dong · Bo Han · Feng Liu
- abstract@[open-review](#): Supervised learning aims to train a classifier under the assumption that training and test data are from the same distribution. To ease the above assumption, researchers have studied a more realistic setting: out-of-distribution (OOD) detection, where test data may come from classes that are unknown during training (i.e., OOD data). Due to the unavailability and diversity of OOD data, good generalization ability is crucial for effective OOD detection algorithms. To study the generalization of OOD detection, in this paper, we investigate the probably approximately correct (PAC) learning theory of OOD detection, which is proposed by researchers as an open problem. First, we find a necessary condition for the learnability of OOD detection. Then, using this condition, we prove several impossibility theorems for the learnability of OOD detection under some scenarios. Although the impossibility theorems are frustrating, we find that some conditions of these impossibility theorems may not hold in some practical scenarios. Based on this observation, we next give several necessary and sufficient conditions to characterize the learnability of OOD detection in some practical scenarios. Lastly, we also offer theoretical supports for several representative OOD detection works based on our OOD theory.

## [Non-stationary Bandits with Knapsacks](#)

- Shang Liu · Jiashuo Jiang · Xiaocheng Li
- abstract@[open-review](#): In this paper, we study the problem of bandits with knapsacks (BwK) in a non-stationary environment. The BwK problem generalizes the multi-arm bandit (MAB) problem to model the resource consumption associated with playing each arm. At each time, the decision maker/player chooses to play an arm, and s/he will receive a reward and consume certain amount of resource from each of the multiple resource types. The objective is to maximize the cumulative reward over a finite horizon subject to some knapsack constraints on the resources. Existing works study the BwK problem under either a stochastic or adversarial environment. Our paper considers a non-stationary environment which continuously interpolates between these two extremes. We first show that the traditional notion of variation budget is insufficient to characterize the non-stationarity of the BwK problem for a sublinear regret due to the presence of the constraints, and then we propose a new notion of global non-stationarity measure. We employ both non-stationarity measures to derive upper and lower bounds for the problem. Our results are based on a primal-dual analysis of the underlying linear programs and highlight the interplay between the constraints and the non-stationarity. Finally, we also extend the non-stationarity measure to the problem of online convex optimization with constraints and obtain new regret bounds accordingly.

## [Alleviating Adversarial Attacks on Variational Autoencoders with MCMC](#)

- Anna Kuzina · Max Welling · Jakub Tomczak
- abstract@[open-review](#): Variational autoencoders (VAEs) are latent variable models that can generate complex objects and provide meaningful latent representations. Moreover, they could be further used in downstream tasks such as classification. As previous work has shown, one can easily fool VAEs to produce unexpected latent representations and reconstructions for a visually slightly modified input. Here, we examine several objective functions for adversarial attacks construction proposed previously and present a solution to alleviate the effect of these attacks. Our method utilizes the Markov Chain Monte Carlo (MCMC) technique in the inference step that we motivate with a theoretical analysis. Thus, we do not incorporate any extra costs during training and the performance on non-attacked inputs is not decreased. We validate our approach on a variety of datasets (MNIST, Fashion MNIST, Color MNIST, CelebA) and VAE configurations (\$\beta\$-VAE, NVAE, \$\beta\$-TCVAE), and show that our approach consistently improves the model robustness to adversarial attacks.

## [Learning Substructure Invariance for Out-of-Distribution Molecular Representations](#)

- Nianzu Yang · Kaipeng Zeng · Qitian Wu · Xiaosong Jia · Junchi Yan
- abstract@[open-review](#): Molecule representation learning (MRL) has been extensively studied and current methods have shown promising power for various tasks, e.g., molecular property prediction and target identification. However, a common hypothesis of existing methods is that either the model development or experimental evaluation is mostly based on i.i.d. data across training and testing. Such a hypothesis can be violated in real-world applications where testing molecules could come from new environments, bringing about serious performance degradation or unexpected prediction. We propose a new representation learning framework entitled MoleOOD to enhance the robustness of MRL models against such distribution shifts, motivated by an observation that the (bio)chemical properties of molecules are usually invariantly associated with certain privileged molecular substructures across different environments (e.g., scaffolds, sizes, etc.). Specifically, We introduce an environment inference model to identify the latent factors that impact data generation from different distributions in a fully data-driven manner. We also propose a new learning objective to guide the molecule encoder to leverage environment-invariant substructures that more stably relate with the labels across environments. Extensive experiments on ten real-world datasets demonstrate that our model has a stronger generalization ability than existing methods under various out-of-distribution (OOD) settings, despite the absence of manual specifications of environments. Particularly, our method achieves up to 5.9% and 3.9% improvement over the strongest baselines on OGB and DrugOOD benchmarks in terms of ROC-AUC, respectively.

## [Learning to Share in Multi-Agent Reinforcement Learning](#)

- Yuxuan Yi · Ge Li · Yaowei Wang · Zongqing Lu
- abstract@[open-review](#): In this paper, we study the problem of networked multi-agent reinforcement learning (MARL), where a number of agents are deployed as a partially connected network and each interacts only with nearby agents. Networked MARL requires all agents to make decisions in a decentralized manner to optimize a global objective with restricted communication between neighbors over the network. Inspired by the fact that sharing plays a key role in human's learning of cooperation, we propose LToS, a hierarchically decentralized MARL framework that enables agents to learn to dynamically share reward with neighbors so as to encourage agents to cooperate on the global objective through collectives. For each agent, the high-level policy learns how to share reward with neighbors to decompose the global objective, while the low-level policy learns to optimize the local objective induced by the high-level policies in the neighborhood. The two policies form a bi-level optimization and learn alternately. We empirically demonstrate that LToS outperforms existing methods in both social dilemma and networked MARL scenarios across scales.

## [Are AlphaZero-like Agents Robust to Adversarial Perturbations?](#)

- Li-Cheng Lan · Huan Zhang · Ti-Rong Wu · Meng-Yu Tsai · I-Chen Wu · Cho-Jui Hsieh
- abstract@[open-review](#): AlphaZero (AZ) has demonstrated that neural network-based Go AIs can surpass human game performance by a large margin. Even without any Monte Carlo tree search (MCTS), the Policy-Value neural networks (PV-NN), which served as a heuristic in AZ, achieve comparable performance to professional players that most humans cannot reach. However, the robustness of those AZ agents and their PV-NNs has not been studied in the literature. Although it is well known that convolutional networks used in computer vision are not robust, existing adversarial attacks cannot be directly applied due to the difficulty of defining semantically invariant perturbations. Further, the discrete nature of Go prevents the use of efficient gradient-based attacks. In this paper, we develop the first adversarial attack on AZ agents of Go. We show that both PV-NNs and AZ agents with few simulations can be fooled by adding one or two irrelevant stones. For example, on 58% of the AlphaGo Zero self-play games, our method can make the widely used KataGo agent with 50 simulations play a losing action by adding two meaningless stones on the board. Moreover, these mistakes are so obvious that even normal humans can independently interpret them. In the experiments, we use the proposed method to examine the robustness of four publicly available Go AZ agents and one NoGo AZ agent. The results show that those agents with few simulations are vulnerable and will make mistakes way below their level.

## [Point-M2AE: Multi-scale Masked Autoencoders for Hierarchical Point Cloud Pre-training](#)

- Renrui Zhang · Ziyu Guo · Peng Gao · Rongyao Fang · Bin Zhao · Dong Wang · Yu Qiao · Hongsheng Li
- abstract@[open-review](#): Masked Autoencoders (MAE) have shown great potentials in self-supervised pre-training for language and 2D image transformers. However, it still remains an open question on how to exploit masked autoencoding for learning representation of irregular 3D point clouds. In this paper, we propose Point-M2AE, a strong Multi-scale MAE pre-training framework for hierarchical self-supervised learning of 3D point clouds. Unlike the standard transformer in MAE, we modify the encoder and decoder into pyramid architectures to progressively model spatial geometries and capture both fine-grained and high-level semantics of 3D shapes. For the encoder that downsamples point tokens by stages, we design a multi-scale masking strategy to generate consistent visible regions across scales and adopt a local spatial self-attention mechanism to focus on neighboring patterns. By multi-scale token propagation, the lightweight decoder gradually upsamples point tokens with complementary skip connections from the encoder, which further promotes the reconstruction from a global-to-local perspective. Extensive experiments demonstrate the state-of-the-art performance of Point-M2AE for 3D representation learning. With a frozen encoder after pre-training, Point-M2AE achieves 92.9% accuracy for linear SVM on ModelNet40, even surpassing some fully trained methods. By fine-tuning on downstream tasks, Point-M2AE achieves 86.43% accuracy on ScanObjectNN, +3.36% to the second-best, and largely benefits the few-shot classification, part segmentation and 3D object detection with the hierarchical learning scheme.

## [Robust Model Selection and Nearly-Proper Learning for GMMs](#)

- Allen Liu · Jerry Li · Ankur Moitra
- abstract@[open-review](#): In learning theory, a standard assumption is that the data is generated from a finite mixture model. But what happens when the number of components is not known in advance? The problem of estimating the number of components, also called model selection, is important in its own right but there are essentially no known efficient algorithms with provable guarantees. In this work, we study the problem of model selection for univariate Gaussian mixture models (GMMs). Given  $\text{poly}(k/\epsilon)$  samples from a distribution that is  $\epsilon$ -close in TV distance to a GMM with  $k$  components, we can construct a GMM with  $\widetilde{O}(k)$  components that approximates the distribution to within  $\widetilde{O}(\epsilon)$  in  $\text{poly}(k/\epsilon)$  time. Thus we are able to approximately determine the minimum number of components needed to fit the distribution within a logarithmic factor. Moreover, by adapting the techniques we obtain similar results for reconstructing Fourier-sparse signals. Prior to our work, the only known algorithms for learning arbitrary univariate GMMs either output significantly more than  $k$  components (e.g.  $k/\epsilon^2$  components for kernel density estimates) or run in time exponential in  $k$ .

## [Can Adversarial Training Be Manipulated By Non-Robust Features?](#)

- Lue Tao · Lei Feng · Hongxin Wei · Jinfeng Yi · Sheng-Jun Huang · Songcan Chen
- abstract@[open-review](#): Adversarial training, originally designed to resist test-time adversarial examples, has shown to be promising in mitigating training-time availability attacks. This defense ability, however, is challenged in this paper. We identify a novel threat model named stability attacks, which aims to hinder robust availability by slightly manipulating the training data. Under this threat, we show that adversarial training using a conventional defense budget  $\epsilon$  provably fails to provide test robustness in a simple statistical setting, where the non-robust features of the training data can be reinforced by  $\epsilon$ -bounded perturbation. Further, we analyze the necessity of enlarging the defense budget to counter stability attacks. Finally, comprehensive experiments demonstrate that stability attacks are harmful on benchmark datasets, and thus the adaptive defense is necessary to maintain robustness.

## [Optimal Transport-based Identity Matching for Identity-invariant Facial Expression Recognition](#)

- Daeha Kim · Byung Cheol Song
- abstract@[open-review](#): Identity-invariant facial expression recognition (FER) has been one of the challenging computer vision tasks. Since conventional FER schemes do not explicitly address the inter-identity variation of facial expressions, their neural network models still operate depending on facial identity. This paper proposes to quantify the inter-identity variation by utilizing pairs of similar expressions explored through a specific matching process. We formulate the identity matching process as an Optimal Transport (OT) problem. Specifically, to find pairs of similar expressions from different identities, we define the inter-feature similarity as a transportation cost. Then, optimal identity matching to find the optimal flow with minimum transportation cost is performed by Sinkhorn-Knopp iteration. The proposed matching method is not only easy to plug in to other models, but also requires only acceptable computational overhead. Extensive simulations prove that the proposed FER method improves the PCC/CCC performance by up to 10% or more compared to the runner-up on wild datasets. The source code and software demo are available at [https://github.com/kdhht2334/ELIM\\_FER](https://github.com/kdhht2334/ELIM_FER).

## [Benign Underfitting of Stochastic Gradient Descent](#)

- Tomer Koren · Roi Livni · Yishay Mansour · Uri Sherman

- abstract@[open-review](#): We study to what extent may stochastic gradient descent (SGD) be understood as a ``conventional'' learning rule that achieves generalization performance by obtaining a good fit to training data. We consider the fundamental stochastic convex optimization framework, where (one pass, \$textit{without}\$\_-replacement) SGD is classically known to minimize the population risk at rate  $\mathcal{O}(1/\sqrt{n})$ , and prove that, surprisingly, there exist problem instances where the SGD solution exhibits both empirical risk and generalization gap of  $\mathcal{O}(1)$ . Consequently, it turns out that SGD is not algorithmically stable in  $\text{any}$  sense, and its generalization ability cannot be explained by uniform convergence or any other currently known generalization bound technique for that matter (other than that of its classical analysis). We then continue to analyze the closely related  $\text{with}$ -replacement SGD, for which we show that an analogous phenomenon does not occur and prove that its population risk does in fact converge at the optimal rate. Finally, we interpret our main results in the context of without-replacement SGD for finite-sum convex optimization problems, and derive upper and lower bounds for the multi-epoch regime that significantly improve upon previously known results.

## [Self-supervised Amodal Video Object Segmentation](#)

- Jian Yao Â· Yuxin Hong Â· Chiyu Wang Â· Tianjun Xiao Â· Tong He Â· Yanwei Fu Â· Francesco Locatello Â· David P Wipf Â· Zheng Zhang
- abstract@[open-review](#): Amodal perception requires inferring the full shape of an object that is partially occluded. This task is particularly challenging on two levels: (1) it requires more information than what is contained in the instant retina or imaging sensor, (2) it is difficult to obtain enough well-annotated amodal labels for supervision. To this end, this paper develops a new framework of Self-supervised amodal Video object segmentation (SaVos). Our method efficiently leverages the visual information of video temporal sequences to infer the amodal mask of objects. The key intuition is that the occluded part of an object can be explained away if that part is visible in other frames, possibly deformed as long as the deformation can be reasonably learned. Accordingly, we derive a novel self-supervised learning paradigm that efficiently utilizes the visible object parts as the supervision to guide the training on videos. In addition to learning type prior to complete masks for known types, SaVos also learns the spatiotemporal prior, which is also useful for the amodal task and could generalize to unseen types. The proposed framework achieves the state-of-the-art performance on the synthetic amodal segmentation benchmark FISHBOWL and the real world benchmark KINS-Video-Car. Further, it lends itself well to being transferred to novel distributions using test-time adaptation, outperforming existing models even after the transfer to a new distribution.

## [MaskPlace: Fast Chip Placement via Reinforced Visual Representation Learning](#)

- Yao Lai Â· Yao Mu Â· Ping Luo
- abstract@[open-review](#): Placement is an essential task in modern chip design, aiming at placing millions of circuit modules on a 2D chip canvas. Unlike the human-centric solution, which requires months of intense effort by hardware engineers to produce a layout to minimize delay and energy consumption, deep reinforcement learning has become an emerging autonomous tool. However, the learning-centric method is still in its early stage, impeded by a massive design space of size ten to the order of a few thousand. This work presents MaskPlace to automatically generate a valid chip layout design within a few hours, whose performance can be superior or comparable to recent advanced approaches. It has several appealing benefits that prior arts do not have. Firstly, MaskPlace recasts placement as a problem of learning pixel-level visual representation to comprehensively describe millions of modules on a chip, enabling placement in a high-resolution canvas and a large action space. It outperforms recent methods that represent a chip as a hypergraph. Secondly, it enables training the policy network by an intuitive reward function with dense reward, rather than a complicated reward function with sparse reward from previous methods. Thirdly, extensive experiments on many public benchmarks show that MaskPlace outperforms existing RL approaches in all key performance metrics, including wirelength, congestion, and density. For example, it achieves 60%-90% wirelength reduction and guarantees zero overlaps. We believe MaskPlace can improve AI-assisted chip layout design. The deliverables are released at <https://laiyao1.github.io/maskplace>.

## [Debugging and Explaining Metric Learning Approaches: An Influence Function Based Perspective](#)

- Ruofan Liu Â· Yun Lin Â· XIANGLIN YANG Â· Jin Song Dong
- abstract@[open-review](#): Deep metric learning (DML) learns a generalizable embedding space where the representations of semantically similar samples are closer. Despite achieving good performance, the state-of-the-art models still suffer from the generalization errors such as farther similar samples and closer dissimilar samples in the space. In this work, we design an empirical influence function (EIF), a debugging and explaining technique for the generalization errors of state-of-the-art metric learning models. EIF is designed to efficiently identify and quantify how a subset of training samples contributes to the generalization errors. Moreover, given a user-specific error, EIF can be used to relabel a potentially noisy training sample as mitigation. In our quantitative experiment, EIF outperforms the traditional baseline in identifying more relevant training samples with statistical significance and 33.5% less time. In the field study on well-known datasets such as CUB200, CARS196, and InShop, EIF identifies 4.4%, 6.6%, and 17.7% labelling mistakes, indicating the direction of the DML community to further improve the model performance. Our code is available at [https://github.com/lindsey98/Influencefunctionmetric\\_learning](https://github.com/lindsey98/Influencefunctionmetric_learning).

## [Nearly-Tight Bounds for Testing Histogram Distributions](#)

- ClÃ©ment L Canonne Â· Ilias Diakonikolas Â· Daniel Kane Â· Sihan Liu
- abstract@[open-review](#): We investigate the problem of testing whether a discrete probability distribution over an ordered domain is a histogram on a specified number of bins. One of the most common tools for the succinct approximation of data,  $k$ -histograms over  $[n]$ , are probability distributions that are piecewise constant over a set of  $k$  intervals. Given samples from an unknown distribution  $\mathbf{p}$  on  $[n]$ , we want to distinguish between the cases that  $\mathbf{p}$  is a  $k$ -histogram versus far from any  $k$ -histogram, in total variation distance. Our main result is a sample near-optimal and computationally efficient algorithm for this testing problem, and a nearly-matching (within logarithmic factors) sample complexity lower bound, showing that the testing problem has sample complexity  $\tilde{\Theta}(\sqrt{nk}/\epsilon^2 + \sqrt{n}/\epsilon^2)$ .

## [Grow and Merge: A Unified Framework for Continuous Categories Discovery](#)

- Xinwei Zhang Â· Jianwen Jiang Â· Yutong Feng Â· Zhi-Fan Wu Â· Xibin Zhao Â· Hai Wan Â· Mingqian Tang Â· Rong Jin Â· Yue Gao
- abstract@[open-review](#): Although a number of studies are devoted to novel category discovery, most of them assume a static setting where both labeled and unlabeled data are given at once for finding new categories. In this work, we focus on the application scenarios where unlabeled data are continuously fed into the category discovery system. We refer to it as the {bf Continuous Category Discovery} ({bf CCD}) problem, which is significantly more challenging than the static setting. A common challenge faced by novel category discovery is that different sets of features are needed for classification and category discovery: class discriminative features are preferred for classification, while rich and diverse features are more suitable for new category mining. This challenge becomes more severe for dynamic setting as the system is asked to deliver good performance for known classes over time, and at the same time continuously discover new classes from unlabeled data. To address this challenge, we develop a framework of {bf Grow and Merge} ({bf GM}) that works by alternating between a growing phase and a merge phase: in the growing phase, it increases the diversity of features through a continuous self-supervised learning for effective category mining, and in the merging phase, it merges the grown model with a static one to ensure satisfying performance for known classes. Our extensive studies verify that the proposed GM framework is significantly more effective than the state-of-the-art approaches for continuous category discovery.

## [Sharp Analysis of Stochastic Optimization under Global Kurdyka-Lojasiewicz Inequality](#)

- Jalal Etesami Â· Ilyas Fatkhullin Â· Niao He Â· Negar Kiyavash
- abstract@[open-review](#): We study the complexity of finding the global solution to stochastic nonconvex optimization when the objective function satisfies global Kurdyka-Lojasiewicz (KL) inequality and the queries from stochastic gradient oracles satisfy mild expected smoothness assumption. We first

introduce a general framework to analyze Stochastic Gradient Descent (SGD) and its associated nonlinear dynamics under the setting. As a byproduct of our analysis, we obtain a sample complexity of  $\mathcal{O}(\epsilon^{-(4-\alpha)/\alpha})$  for SGD when the objective satisfies the so called  $\alpha$ -P(L) condition, where  $\alpha$  is the degree of gradient domination. Furthermore, we show that a modified SGD with variance reduction and restarting (PAGER) achieves an improved sample complexity of  $\mathcal{O}(\epsilon^{-2/\alpha})$  when the objective satisfies the average smoothness assumption. This leads to the first optimal algorithm for the important case of  $\alpha=1$  which appears in applications such as policy optimization in reinforcement learning.

## [Private Graph All-Pairwise-Shortest-Path Distance Release with Improved Error Rate](#)

- Chenglin Fan · Ping Li · Xiaoyun Li
- abstract@[open-review](#): Releasing all pairwise shortest path (APSP) distances between vertices on general graphs under weight Differential Privacy (DP) is known as a challenging task. In previous work, to achieve DP with some fixed budget, with high probability the maximal absolute error among all published pairwise distances is roughly  $O(n)$  where  $n$  is the number of nodes. It was shown that this error could be reduced for some special graphs, which, however, is hard for general graphs. Therefore, whether the approximation error can be reduced to sublinear is posted as an interesting open problem. In this paper, we break the linear barrier on the distance approximation error of previous result, by proposing an algorithm that releases a constructed synthetic graph privately. Computing all pairwise distances on the constructed graph only introduces  $O(n^{1/2})$  error in answering all pairwise shortest path distances for fixed privacy parameter. Our method is based on a novel graph diameter (link length) augmentation via constructing ``shortcuts'' for the paths. By adding a set of shortcut edges to the original graph, we show that any node pair has a shortest path with link length  $O(n^{1/2})$ . Then by adding noises with some positive mean to the edge weights, the new graph is differentially private and can be published to answer all pairwise shortest path distances with  $O(n^{1/2})$  approximation error using standard APSP computation. Numerical examples are also provided. Additionally, we also consider the graph with small feedback vertex set number. A feedback vertex set (FVS) of a graph is a set of vertices whose removal leaves a graph without cycles, and the feedback vertex set number of a graph,  $k$ , is the size of a smallest feedback vertex set. We propose a DP algorithm with error rate  $O(k)$ , which improves the error of general graphs provided  $k=o(n^{1/2})$ .

## [Learning Manifold Dimensions with Conditional Variational Autoencoders](#)

- Yijia Zheng · Tong He · Yixuan Qiu · David P Wipf
- abstract@[open-review](#): Although the variational autoencoder (VAE) and its conditional extension (CVAE) are capable of state-of-the-art results across multiple domains, their precise behavior is still not fully understood, particularly in the context of data (like images) that lie on or near a low-dimensional manifold. For example, while prior work has suggested that the globally optimal VAE solution can learn the correct manifold dimension, a necessary (but not sufficient) condition for producing samples from the true data distribution, this has never been rigorously proven. Moreover, it remains unclear how such considerations would change when various types of conditioning variables are introduced, or when the data support is extended to a union of manifolds (e.g., as is likely the case for MNIST digits and related). In this work, we address these points by first proving that VAE global minima are indeed capable of recovering the correct manifold dimension. We then extend this result to more general CVAEs, demonstrating practical scenarios whereby the conditioning variables allow the model to adaptively learn manifolds of varying dimension across samples. Our analyses are also supported by numerical results on both synthetic and real-world datasets.

## [Privacy of Noisy Stochastic Gradient Descent: More Iterations without More Privacy Loss](#)

- Jason Altschuler · Kunal Talwar
- abstract@[open-review](#): A central issue in machine learning is how to train models on sensitive user data. Industry has widely adopted a simple algorithm: Stochastic Gradient Descent with noise (a.k.a. Stochastic Gradient Langevin Dynamics). However, foundational theoretical questions about this algorithm's privacy loss remain open---even in the seemingly simple setting of smooth convex losses over a bounded domain. Our main result resolves these questions: for a large range of parameters, we characterize the differential privacy up to a constant. This result reveals that all previous analyses for this setting have the wrong qualitative behavior. Specifically, while previous privacy analyses increase ad infinitum in the number of iterations, we show that after a small burn-in period, running SGD longer leaks no further privacy. Our analysis departs completely from previous approaches based on fast mixing, instead using techniques based on optimal transport (namely, Privacy Amplification by Iteration) and the sampled Gaussian mechanism (namely, Privacy Amplification by Sampling). Our techniques readily extend to other settings, e.g., strongly convex losses, non-uniform stepsizes, arbitrary batch sizes, and random or cyclic batches.

## [Annihilation of Families of Spurious Minima in Two-Layer ReLU Networks](#)

- Yossi Arjevani · Michael Field
- abstract@[open-review](#): We study the optimization problem associated with fitting two-layer ReLU neural networks with respect to the squared loss, where labels are generated by a target network. Use is made of the rich symmetry structure to develop a novel set of tools for studying the mechanism by which over-parameterization annihilates spurious minima through. Sharp analytic estimates are obtained for the loss and the Hessian spectrum at different minima, and it is shown that adding neurons can turn symmetric spurious minima into saddles through a local mechanism that does not generate new spurious minima; minima of smaller symmetry require more neurons. Using Cauchy's interlacing theorem, we prove the existence of descent directions in certain subspaces arising from the symmetry structure of the loss function. This analytic approach uses techniques, new to the field, from algebraic geometry, representation theory and symmetry breaking, and confirms rigorously the effectiveness of over-parameterization in making the associated loss landscape accessible to gradient-based methods. For a fixed number of neurons and inputs, the spectral results remain true under symmetry breaking perturbation of the target.

## [On Margins and Generalisation for Voting Classifiers](#)

- Felix Biggs · Valentina Zantedeschi · Benjamin Guedj
- abstract@[open-review](#): We study the generalisation properties of majority voting on finite ensembles of classifiers, proving margin-based generalisation bounds via the PAC-Bayes theory. These provide state-of-the-art guarantees on a number of classification tasks. Our central results leverage the Dirichlet posteriors studied recently by Zantedeschi et al. (2021) for training voting classifiers; in contrast to that work our bounds apply to non-randomised votes via the use of margins. Our contributions add perspective to the debate on the ``margins theory'' proposed by Schapire et al. (1998) for the generalisation of ensemble classifiers.

## [Simulated User Studies for Explanation Evaluation](#)

- Valerie Chen · Nari Johnson · Nicholay Topin · Gregory Plumb · Ameet Talwalkar
- abstract@[open-review](#): A growing body of research runs human subject evaluations to study whether providing users with explanations of machine learning models can help them with practical real-world use cases. However, running user studies is challenging and costly, and consequently each study typically only evaluates a limited number of different settings, e.g., studies often only evaluate a few arbitrarily selected model explanation methods. To address these challenges and aid user study design, we introduce Simulated Evaluations (SimEvals). SimEvals involve training algorithmic agents that take as input the information content (such as model explanations) that would be presented to the user, to predict answers to the use case of interest. The algorithmic agent's test set accuracy provides a measure of the predictiveness of the information content for the downstream use case. We run a comprehensive evaluation on three real-world use cases (forward simulation, model debugging, and counterfactual reasoning) to demonstrate that

SimEvals can effectively identify which explanation methods will help humans for each use case. These results provide evidence that `\simevals{}` can be used to efficiently screen an important set of user study design decisions, e.g., selecting which explanations should be presented to the user, before running a potentially costly user study.

## [Earthformer: Exploring Space-Time Transformers for Earth System Forecasting](#)

- Zhihan Gao · Xingjian Shi · Hao Wang · Yi Zhu · Yuyang (Bernie) Wang · Mu Li · Dit-Yan Yeung
- abstract@[open-review](#): It has been studied for centuries to predict the evolution of the Earth system due to its significant impact on human lives. Conventionally, Earth system (e.g., weather and climate) forecasting models rely on numerical simulation of complex physical models and are hence expensive in both computational resources and domain expertise. With the explosive growth of Earth observation data in the past decade, data-driven models that apply Deep Learning (DL) are demonstrating impressive potential for various Earth system forecasting tasks. So far, these DL models mainly use Convolutional Neural Networks (CNNs) or Recurrent Neural Networks (RNNs) as the basic building blocks. The Transformer architecture, despite its broad success in other domains, has limited adoption for Earth system forecasting. In this paper, we propose `\emph{Earthformer}`, a space-time Transformer for Earth system forecasting. Earthformer is based on a generic, flexible and efficient space-time attention block, named `\emph{Cuboid Attention}`, which decomposes the data to cuboids and applies cuboid-level self-attention in parallel. These cuboids are further connected with a collection of global vectors. We conduct experiments on the MovingMNIST dataset and a newly proposed chaotic N-body MNIST dataset to verify the effectiveness of cuboid attention and figure out the best design for Earthformer. Experiments on two real-world benchmarks about precipitation nowcasting and El Niño/Southern Oscillation (ENSO) forecasting show Earthformer achieves state-of-the-art performance.

## [Get More at Once: Alternating Sparse Training with Gradient Correction](#)

- Li Yang · Jian Meng · Jae-sun Seo · Deliang Fan
- abstract@[open-review](#): Recently, a new trend of exploring training sparsity has emerged, which remove parameters during training, leading to both training and inference efficiency improvement. This line of works primarily aims to obtain a single sparse model under a pre-defined large sparsity ratio. It leads to a static/fixed sparse inference model that is not capable of adjusting or re-configuring its computation complexity (i.e., inference structure, latency) after training for real-world varying and dynamic hardware resource availability. To enable such run-time or post-training network morphing, the concept of dynamic inference' ortraining-once-for-all' has been proposed to train a single network consisting of multiple sub-nets once, but each sub-net could perform the same inference function with different computing complexity. However, the traditional dynamic inference training method requires a joint training scheme with multi-objective optimization, which suffers from very large training overhead. In this work, for the first time, we propose a novel alternating sparse training (AST) scheme to train multiple sparse sub-nets for dynamic inference without extra training cost compared to the case of training a single sparse model from scratch. Furthermore, to mitigate the interference of weight update among sub-nets, we propose gradient correction within the inner-group iterations to reduce their weight update interference. We validate the proposed AST on multiple datasets against state-of-the-art sparse training method, which shows that AST achieves similar or better accuracy, but only needs to train once to get multiple sparse sub-nets with different sparsity ratios. More importantly, comparing with the traditional joint training based dynamic inference training methodology, the large training overhead is completely eliminated without affecting the accuracy of each sub-net.

## [Local Bayesian optimization via maximizing probability of descent](#)

- Quan Nguyen · Kaiwen Wu · Jacob Gardner · Roman Garnett
- abstract@[open-review](#): Local optimization presents a promising approach to expensive, high-dimensional black-box optimization by sidestepping the need to globally explore the search space. For objective functions whose gradient cannot be evaluated directly, Bayesian optimization presents one promising approach -- we construct a Gaussian process (GP) model of the objective, design a policy to learn about the gradient at the current location through nearby observations of the objective, and use the resulting information to navigate the objective landscape. Previous work has realized this scheme by maximizing the information gained about the gradient, then moving in the direction of the expected gradient. In this paper, we reexamine and refine this approach. We demonstrate that, surprisingly, the expected value of the gradient is not always the direction maximizing the probability of descent, and in fact, these directions may be nearly orthogonal. This observation then inspires an elegant optimization scheme seeking to maximize the probability of descent while moving in the direction of most-likely descent. Experiments on both synthetic and real-world objectives show that our method outperforms previous realizations of this optimization scheme and is competitive against other, significantly more-complex baselines.

## [Can Hybrid Geometric Scattering Networks Help Solve the Maximal Clique Problem?](#)

- Yimeng Min · Frederik Wenkel · Michael Perlmutter · Guy Wolf
- abstract@[open-review](#): We propose a geometric scattering-based graph neural network (GNN) for approximating solutions of the NP-hard maximum clique (MC) problem. We construct a loss function with two terms, one which encourages the network to find a large set of nodes and the other which acts as a surrogate for the constraint that the nodes form a clique. We then use this loss to train an efficient GNN architecture that outputs a vector representing the probability for each node to be part of the MC and apply a rule-based decoder to make our final prediction. The incorporation of the scattering transform alleviates the so-called oversmoothing problem that is often encountered in GNNs and would degrade the performance of our proposed setup. Our empirical results demonstrate that our method outperforms representative GNN baselines in terms of solution accuracy and inference speed as well as conventional solvers like GUROBI with limited time budgets. Furthermore, our scattering model builds efficient representation and only consists of  $\sim 0.1\%$  of parameter counts of previous GNN baseline model.

## [projUNN: efficient method for training deep networks with unitary matrices](#)

- Bobak Kiani · Randall Balestrieri · Yann LeCun · Seth Lloyd
- abstract@[open-review](#): In learning with recurrent or very deep feed-forward networks, employing unitary matrices in each layer can be very effective at maintaining long-range stability. However, restricting network parameters to be unitary typically comes at the cost of expensive parameterizations or increased training runtime. We propose instead an efficient method based on rank-\$k\$ updates -- or their rank-\$k\$ approximation -- that maintains performance at a nearly optimal training runtime. We introduce two variants of this method, named Direct (projUNN-D) and Tangent (projUNN-T) projected Unitary Neural Networks, that can parameterize full \$N\$-dimensional unitary or orthogonal matrices with a training runtime scaling as  $O(kN^2)$ . Our method either projects low-rank gradients onto the closest unitary matrix (projUNN-T) or transports unitary matrices in the direction of the low-rank gradient (projUNN-D). Even in the fastest setting ( $k=1$ ), projUNN is able to train a model's unitary parameters to reach comparable performances against baseline implementations. In recurrent neural network settings, projUNN closely matches or exceeds benchmarked results from prior unitary neural networks. Finally, we preliminarily explore projUNN in training orthogonal convolutional neural networks, which are currently unable to outperform state of the art models but can potentially enhance stability and robustness at large depth.

## [Data-IQ: Characterizing subgroups with heterogeneous outcomes in tabular data](#)

- Nabeel Seedat · Jonathan Crabb · Ioana Bica · Mihaela van der Schaar
- abstract@[open-review](#): High model performance, on average, can hide that models may systematically underperform on subgroups of the data. We consider the tabular setting, which surfaces the unique issue of outcome heterogeneity - this is prevalent in areas such as healthcare, where patients with similar features can have different outcomes, thus making reliable predictions challenging. To tackle this, we propose Data-IQ, a framework to systematically stratify examples into subgroups with respect to their outcomes. We do this by analyzing the behavior of individual examples during

training, based on their predictive confidence and, importantly, the aleatoric (data) uncertainty. Capturing the aleatoric uncertainty permits a principled characterization and then subsequent stratification of data examples into three distinct subgroups (Easy, Ambiguous, Hard). We experimentally demonstrate the benefits of Data-IQ on four real-world medical datasets. We show that Data-IQ's characterization of examples is most robust to variation across similarly performant (yet different models), compared to baselines. Since Data-IQ can be used with any ML model (including neural networks, gradient boosting etc.), this property ensures consistency of data characterization, while allowing flexible model selection. Taking this a step further, we demonstrate that the subgroups enable us to construct new approaches to both feature acquisition and dataset selection. Furthermore, we highlight how the subgroups can inform reliable model usage, noting the significant impact of the Ambiguous subgroup on model generalization.

## [Improved Algorithms for Neural Active Learning](#)

- Yikun Ban · Yuheng Zhang · Hanghang Tong · Arindam Banerjee · Jingrui He
- abstract@[open-review](#): We improve the theoretical and empirical performance of neural-network(NN)-based active learning algorithms for the non-parametric streaming setting. In particular, we introduce two regret metrics by minimizing the population loss that are more suitable in active learning than the one used in state-of-the-art (SOTA) related work. Then, the proposed algorithm leverages the powerful representation of NNs for both exploitation and exploration, has the query decision-maker tailored for  $k$ -class classification problems with the performance guarantee, utilizes the full feedback, and updates parameters in a more practical and efficient manner. These careful designs lead to a better regret upper bound, improving by a multiplicative factor  $\mathcal{O}(\log T)$  and removing the curse of both input dimensionality and the complexity of the function to be learned. Furthermore, we show that the algorithm can achieve the same performance as the Bayes-optimal classifier in the long run under the hard-margin setting in classification problems. In the end, we use extensive experiments to evaluate the proposed algorithm and SOTA baselines, to show the improved empirical performance.

## [Distributionally robust weighted k-nearest neighbors](#)

- Shixiang Zhu · Liyan Xie · Minghe Zhang · Rui Gao · Yao Xie
- abstract@[open-review](#): Learning a robust classifier from a few samples remains a key challenge in machine learning. A major thrust of research has been focused on developing k-nearest neighbor (k-NN) based algorithms combined with metric learning that captures similarities between samples. When the samples are limited, robustness is especially crucial to ensure the generalization capability of the classifier. In this paper, we study a minimax distributionally robust formulation of weighted k-nearest neighbors, which aims to find the optimal weighted k-NN classifiers that hedge against feature uncertainties. We develop an algorithm, Dr.k-NN, that efficiently solves this functional optimization problem and features in assigning minimax optimal weights to training samples when performing classification. These weights are class-dependent, and are determined by the similarities of sample features under the least favorable scenarios. When the size of the uncertainty set is properly tuned, the robust classifier has a smaller Lipschitz norm than the vanilla k-NN, and thus improves the generalization capability. We also couple our framework with neural-network-based feature embedding. We demonstrate the competitive performance of our algorithm compared to the state-of-the-art in the few-training-sample setting with various real-data experiments.

## [An Algorithm for Learning Switched Linear Dynamics from Data](#)

- Guillaume Berger · Monal Narasimhamurthy · Kandai Watanabe · Morteza Lahijanian · Sriram Sankaranarayanan
- abstract@[open-review](#): In this paper, we present an algorithm for learning switched linear dynamical systems in discrete-time from data that may include noisy observations of the full system state or output observations. Switched linear systems generalize linear systems by using multiple linear dynamical modes to explain the data within some desired tolerance. They arise quite naturally in many applications such as robotics and cyber-physical systems. Learning these dynamics from given data is a NP-hard problem that is very closely related to the  $k$ -linear regression problem of fitting  $k > 1$  linear models to given data. A direct  $\mathcal{O}(n^k)$  mixed integer linear programming approach yields time complexity that is exponential in the number of data points. In this paper, we modify the problem formulation slightly to yield an algorithm that is linear in the size of the data while being exponential in the number of state variables and the desired number of modes. To do so, we combine classic ideas from the ellipsoidal method for solving convex optimization problems, and well-known oracle separation results in non-smooth optimization. We demonstrate our approach on a set of micro-benchmarks and a few interesting real-world data sets. Our evaluation on a prototype implementation suggests that the benefits of this algorithm can be made practical even against highly optimized commercial mixed-integer solvers.

## [Sparse Winning Tickets are Data-Efficient Image Recognizers](#)

- Mukund Varma T · Xuxi Chen · Zhenyu Zhang · Tianlong Chen · Subhashini Venugopalan · Zhangyang Wang
- abstract@[open-review](#): Improving performance of deep networks in data limited regimes has warranted much attention. In this work, we show that "winning tickets" (small sub-networks) obtained via magnitude pruning based on the lottery ticket hypothesis (Frankle & Carbin, 2018), apart from being sparse are also effective recognizers in data limited regimes. Based on extensive experiments, we find that in low data regimes (datasets of 50-100 examples per class), sparse winning tickets substantially outperform the original dense networks. This approach, when combined with augmentations or fine-tuning from a self-supervised backbone network, shows further improvements in performance by as much as 16% (absolute) on low sample datasets and long-tailed classification. Further, sparse winning tickets are more robust to synthetic noise and distribution shifts compared to their dense counterparts. Our analysis of winning tickets on small datasets indicates that, though sparse, the networks retain density in the initial layers and their representations are more generalizable. Code will be made available after acceptance.

## [Sobolev Acceleration and Statistical Optimality for Learning Elliptic Equations via Gradient Descent](#)

- Yiping Lu · Jose Blanchet · Lexing Ying
- abstract@[open-review](#): In this paper, we study the statistical limits in terms of Sobolev norms of gradient descent for solving inverse problem from randomly sampled noisy observations using a general class of objective functions. Our class of objective functions includes Sobolev training for kernel regression, Deep Ritz Methods (DRM), and Physics Informed Neural Networks (PINN) for solving elliptic partial differential equations (PDEs) as special cases. We consider a potentially infinite-dimensional parameterization of our model using a suitable Reproducing Kernel Hilbert Space and a continuous parameterization of problem hardness through the definition of kernel integral operators. We prove that gradient descent over this objective function can also achieve statistical optimality and the optimal number of passes over the data increases with sample size. Based on our theory, we explain an implicit acceleration of using a Sobolev norm as the objective function for training, inferring that the optimal number of epochs of DRM becomes larger than the number of PINN when both the data size and the hardness of tasks increase, although both DRM and PINN can achieve statistical optimality.

## [LiteTransformerSearch: Training-free On-device Search for Efficient Autoregressive Language Models](#)

- Mojan Javaheripi · Gustavo de Rosa · Subhabrata Mukherjee · Shital Shah · Tomasz Religa · Caio Cesar Teodoro Mendes · Sébastien Bubeck · Farinaz Koushanfar · Debadatta Dey
- abstract@[open-review](#): The Transformer architecture is ubiquitously used as the building block of large-scale autoregressive language models. However, finding architectures with the optimal trade-off between task performance (perplexity) and hardware constraints like peak memory utilization and latency is non-trivial. This is exacerbated by the proliferation of various hardware. We leverage the somewhat surprising empirical observation that the number of decoder parameters in autoregressive Transformers has a high rank correlation with task performance, irrespective of the architecture topology. This observation organically induces a simple search algorithm that uses decoder parameters as a proxy for perplexity without need for any model training. The search phase of our training-free algorithm, dubbed Lightweight Transformer Search (LTS), can be run directly on target devices since it does not require

GPUs. We focus on extracting the pareto-frontier of perplexity versus any hardware performance cost such as latency and/or memory, using on-target-device measurements. We evaluate LTS on diverse devices from ARM CPUs to NVIDIA GPUs and two popular autoregressive Transformer backbones: GPT-2 and Transformer-XL. Results show that the perplexity of 16-layer GPT-2 and Transformer-XL can be achieved with up to  $1.6\text{\AA}$ ,  $2.5\text{\AA}$  faster runtime and  $1.3\text{\AA}$ ,  $2.0\text{\AA}$  lower peak memory utilization. LTS extracts the pareto-frontier in under 3 hours, running on a commodity laptop. We effectively remove the carbon footprint of training during search for hundreds of GPU hours, offering a strong simple baseline for future NAS methods in autoregressive language modeling.

## [Easier Linear Algebra for Distance Matrices](#)

- Piotr Indyk · Sandeep Silwal
- abstract@[open-review](#): The distance matrix of a dataset  $X$  of  $n$  points with respect to a distance function  $f$  represents all pairwise distances between points in  $X$  induced by  $f$ . Due to their wide applicability, distance matrices and related families of matrices have been the focus of many recent algorithmic works. We continue this line of research and take a broad view of algorithm design for distance matrices with the goal of designing fast algorithms, which are specifically tailored for distance matrices, for fundamental linear algebraic primitives. Our results include efficient algorithms for computing matrix-vector products for a wide class of distance matrices, such as the  $\ell_1$  metric for which we get a linear runtime, as well as an  $\Omega(n^2)$  lower bound for any algorithm which computes a matrix-vector product for the  $\ell_\infty$  case, showing a separation between the  $\ell_1$  and the  $\ell_\infty$  metrics. Our upper bound results in conjunction with recent works on the matrix-vector query model have many further downstream applications, including the fastest algorithm for computing a relative error low-rank approximation for the distance matrix induced by  $\ell_1$  and  $\ell_2^2$  functions and the fastest algorithm for computing an additive error low-rank approximation for the  $\ell_2$  metric, in addition to applications for fast matrix multiplication among others. We also give algorithms for constructing distance matrices and show that one can construct an approximate  $\ell_2$  distance matrix in time faster than the bound implied by the Johnson-Lindenstrauss lemma.

## [\(Optimal\) Online Bipartite Matching with Predicted Degrees](#)

- Anders Aamand · Justin Chen · Piotr Indyk
- abstract@[open-review](#): We propose a model for online graph problems where algorithms are given access to an oracle that predicts (e.g., based on past data) the degrees of nodes in the graph. Within this model, we study the classic problem of online bipartite matching, and a natural greedy matching algorithm called MinPredictedDegree, which uses predictions of the degrees of offline nodes. For the bipartite version of a stochastic graph model due to Chung, Lu, and Vu where the expected values of the offline degrees are known and used as predictions, we show that MinPredictedDegree stochastically dominates any other online algorithm, i.e., it is optimal for graphs drawn from this model. Since the "symmetric" version of the model, where all online nodes are identical, is a special case of the well-studied "known i.i.d. model", it follows that the competitive ratio of MinPredictedDegree on such inputs is at least 0.7299. For the special case of graphs with power law degree distributions, we show that MinPredictedDegree frequently produces matchings almost as large as the true maximum matching on such graphs. We complement these results with an extensive empirical evaluation showing that MinPredictedDegree compares favorably to state-of-the-art online algorithms for online matching.

## [Exponentially Improving the Complexity of Simulating the Weisfeiler-Lehman Test with Graph Neural Networks](#)

- Anders Aamand · Justin Chen · Piotr Indyk · Shyam Narayanan · Ronitt Rubinfeld · Nicholas Schiefer · Sandeep Silwal · Tal Wagner
- abstract@[open-review](#): Recent work shows that the expressive power of Graph Neural Networks (GNNs) in distinguishing non-isomorphic graphs is exactly the same as that of the Weisfeiler-Lehman (WL) graph test. In particular, they show that the WL test can be simulated by GNNs. However, those simulations involve neural networks for the  $\text{combine}$  function of size polynomial or even exponential in the number of graph nodes  $n$ , as well as feature vectors of length linear in  $n$ . We present an improved simulation of the WL test on GNNs with exponentially lower complexity. In particular, the neural network implementing the  $\text{combine}$  function in each node has only  $\mathcal{O}(\text{polylog}(n))$  parameters, and the feature vectors exchanged by the nodes of GNN consists of only  $O(\log n)$  bits. We also give logarithmic lower bounds for the feature vector length and the size of the neural networks, showing the (near)-optimality of our construction.

## [Efficient Sampling on Riemannian Manifolds via Langevin MCMC](#)

- Xiang Cheng · Jingzhao Zhang · Suvrit Sra
- abstract@[open-review](#): We study the task of efficiently sampling from a Gibbs distribution  $d\pi = e^{-h} d\text{vol}_g$  over a Riemannian manifold  $M$  via (geometric) Langevin MCMC; this algorithm involves computing exponential maps in random Gaussian directions and is efficiently implementable in practice. The key to our analysis of Langevin MCMC is a bound on the discretization error of the geometric Euler-Murayama scheme, assuming  $\nabla h$  is Lipschitz and  $M$  has bounded sectional curvature. Our error bound matches the error of Euclidean Euler-Murayama in terms of its stepsize dependence. Combined with a contraction guarantee for the geometric Langevin Diffusion under Kendall-Cranston coupling, we prove that the Langevin MCMC iterates lie within  $\tilde{O}(\epsilon)$ -Wasserstein distance of  $\pi$  after  $\tilde{O}(\epsilon^{-2})$  steps, which matches the iteration complexity for Euclidean Langevin MCMC. Our results apply in general settings where  $h$  can be nonconvex and  $M$  can have negative Ricci curvature. Under additional assumptions that the Riemannian curvature tensor has bounded derivatives, and that  $\nabla h$  satisfies a  $CD(c, \infty)$  condition, we analyze the stochastic gradient version of Langevin MCMC, and bound its iteration complexity by  $\tilde{O}(\epsilon^{-2})$  as well.

## [ATD: Augmenting CP Tensor Decomposition by Self Supervision](#)

- Chaoqi Yang · Cheng Qian · Navjot Singh · Cao (Danica) Xiao · M Westover · Edgar Solomonik · Jimeng Sun
- abstract@[open-review](#): Tensor decompositions are powerful tools for dimensionality reduction and feature interpretation of multidimensional data such as signals. Existing tensor decomposition objectives (e.g., Frobenius norm) are designed for fitting raw data under statistical assumptions, which may not align with downstream classification tasks. In practice, raw input tensor can contain irrelevant information while data augmentation techniques may be used to smooth out class-irrelevant noise in samples. This paper addresses the above challenges by proposing augmented tensor decomposition (ATD), which effectively incorporates data augmentations and self-supervised learning (SSL) to boost downstream classification. To address the non-convexity of the new augmented objective, we develop an iterative method that enables the optimization to follow an alternating least squares (ALS) fashion. We evaluate our proposed ATD on multiple datasets. It can achieve 0.8%~2.5% accuracy gain over tensor-based baselines. Also, our ATD model shows comparable or better performance (e.g., up to 15% in accuracy) over self-supervised and autoencoder baselines while using less than 5% of learnable parameters of these baseline models.

## [Residual Multiplicative Filter Networks for Multiscale Reconstruction](#)

- Shayan Shekarforoush · David Lindell · Marcus Brubaker · David Fleet
- abstract@[open-review](#): Coordinate networks like Multiplicative Filter Networks (MFNs) and BACON offer some control over the frequency spectrum used to represent continuous signals such as images or 3D volumes. Yet, they are not readily applicable to problems for which coarse-to-fine estimation is required, including various inverse problems in which coarse-to-fine optimization plays a key role in avoiding poor local minima. We introduce a new coordinate network architecture and training scheme that enables coarse-to-fine optimization with fine-grained control over the frequency support of learned reconstructions. This is achieved with two key innovations. First, we incorporate skip connections so that structure at one scale is preserved when fitting finer-scale structure. Second, we propose a novel initialization scheme to provide control over the model frequency spectrum at each stage of optimization. We demonstrate how these modifications enable multiscale optimization for coarse-to-fine fitting to natural images. We then evaluate our

model on synthetically generated datasets for the the problem of single-particle cryo-EM reconstruction. We learn high resolution multiscale structures, on par with the state-of-the art.

## [Leveraging Factored Action Spaces for Efficient Offline Reinforcement Learning in Healthcare](#)

- Shengpu Tang · Maggie Makar · Michael Sjoding · Finale Doshi-Velez · Jenna Wiens
- abstract@[open-review](#): Many reinforcement learning (RL) applications have combinatorial action spaces, where each action is a composition of sub-actions. A standard RL approach ignores this inherent factorization structure, resulting in a potential failure to make meaningful inferences about rarely observed sub-action combinations; this is particularly problematic for offline settings, where data may be limited. In this work, we propose a form of linear Q-function decomposition induced by factored action spaces. We study the theoretical properties of our approach, identifying scenarios where it is guaranteed to lead to zero bias when used to approximate the Q-function. Outside the regimes with theoretical guarantees, we show that our approach can still be useful because it leads to better sample efficiency without necessarily sacrificing policy optimality, allowing us to achieve a better bias-variance trade-off. Across several offline RL problems using simulators and real-world datasets motivated by healthcare problems, we demonstrate that incorporating factored action spaces into value-based RL can result in better-performing policies. Our approach can help an agent make more accurate inferences within under-explored regions of the state-action space when applying RL to observational datasets.

## [CryptoGCN: Fast and Scalable Homomorphically Encrypted Graph Convolutional Network Inference](#)

- Ran Ran · Wei Wang · Quan Gang · Jieming Yin · Nuo Xu · Wujie Wen
- abstract@[open-review](#): Recently cloud-based graph convolutional network (GCN) has demonstrated great success and potential in many privacy-sensitive applications such as personal healthcare and financial systems. Despite its high inference accuracy and performance on cloud, maintaining data privacy in GCN inference, which is of paramount importance to these practical applications, remains largely unexplored. In this paper, we take an initial attempt towards this and develop CryptoGCN--a homomorphic encryption (HE) based GCN inference framework. A key to the success of our approach is to reduce the tremendous computational overhead for HE operations, which can be orders of magnitude higher than its counterparts in the plaintext space. To this end, we develop an approach that can effectively take advantage of the sparsity of matrix operations in GCN inference to significantly reduce the computational overhead. Specifically, we propose a novel AMA data formatting method and associated spatial convolution methods, which can exploit the complex graph structure and perform efficient matrix-matrix multiplication in HE computation and thus greatly reduce the HE operations. We also develop a co-optimization framework that can explore the trade offs among the accuracy, security level, and computational overhead by judicious pruning and polynomial approximation of activation module in GCNs. Based on the NTU-XVIEW skeleton joint dataset, i.e., the largest dataset evaluated homomorphically by far as we are aware of, our experimental results demonstrate that CryptoGCN outperforms state-of-the-art solutions in terms of the latency and number of homomorphic operations, i.e., achieving as much as a 3.10\$\$\times\$\$ speedup on latency and reduced 77.4% of total Homomorphic Operation Count with a small accuracy loss of 1-1.5%.

## [On Translation and Reconstruction Guarantees of the Cycle-Consistent Generative Adversarial Networks](#)

- Anish Chakrabarty · Swagatam Das
- abstract@[open-review](#): The task of unpaired image-to-image translation has witnessed a revolution with the introduction of the cycle-consistency loss to Generative Adversarial Networks (GANs). Numerous variants, with Cycle-Consistent Adversarial Network (CycleGAN) at their forefront, have shown remarkable empirical performance. The involvement of two unalike data spaces and the existence of multiple solution maps between them are some of the facets that make such architectures unique. In this study, we investigate the statistical properties of such unpaired data translator networks between distinct spaces, bearing the additional responsibility of cycle-consistency. In a density estimation setup, we derive sharp non-asymptotic bounds on the translation errors under suitably characterized models. This, in turn, points out sufficient regularity conditions that maps must obey to carry out successful translations. We further show that cycle-consistency is achieved as a consequence of the data being successfully generated in each space based on observations from the other. In a first-of-its-kind attempt, we also provide deterministic bounds on the cumulative reconstruction error. In the process, we establish tolerable upper bounds on the discrepancy responsible for ill-posedness in such networks.

## [Using Partial Monotonicity in Submodular Maximization](#)

- Loay Mualem · Moran Feldman
- abstract@[open-review](#): Over the last two decades, submodular function maximization has been the workhorse of many discrete optimization problems in machine learning applications. Traditionally, the study of submodular functions was based on \emph{binary} function properties. However, such properties have an inherit weakness, namely, if an algorithm assumes functions that have a particular property, then it provides no guarantee for functions that violate this property, even when the violation is very slight. Therefore, recent works began to consider \emph{continuous} versions of function properties. Probably the most significant among these (so far) are the submodularity ratio and the curvature, which were studied extensively together and separately. The monotonicity property of set functions plays a central role in submodular maximization. Nevertheless, and despite all the above works, no continuous version of this property has been suggested to date (as far as we know). This is unfortunate since submodular functions that are almost monotone often arise in machine learning applications. In this work we fill this gap by defining the \emph{monotonicity ratio}, which is a continuous version of the monotonicity property. We then show that for many standard submodular maximization algorithms one can prove new approximation guarantees that depend on the monotonicity ratio; leading to improved approximation ratios for the common machine learning applications of movie recommendation, quadratic programming and image summarization.

## [Better SGD using Second-order Momentum](#)

- Hoang Tran · Ashok Cutkosky
- abstract@[open-review](#): We develop a new algorithm for non-convex stochastic optimization that finds an  $\$\\epsilon$ -critical point in the optimal  $\$O(\\epsilon^{-3})$  stochastic gradient and Hessian-vector product computations. Our algorithm uses Hessian-vector products to "correct" a bias term in the momentum of SGD with momentum. This leads to better gradient estimates in a manner analogous to variance reduction methods. In contrast to prior work, we do not require excessively large batch sizes and are able to provide an adaptive algorithm whose convergence rate automatically improves with decreasing variance in the gradient estimates. We validate our results on a variety of large-scale deep learning architectures and benchmarks tasks.

## [Differentially Private Online-to-batch for Smooth Losses](#)

- Qinzi Zhang · Ashok Cutkosky · Hoang Tran
- abstract@[open-review](#): We develop a new reduction that converts any online convex optimization algorithm suffering  $\$O(\\sqrt{T})$  regret into an  $\$\\epsilon$ -differentially private stochastic convex optimization algorithm with the optimal convergence rate  $\$\\tilde{O}(1/\\sqrt{T} + 1/\\epsilon T)$  on smooth losses in linear time, forming a direct analogy to the classical non-private ``online-to-batch'' conversion. By applying our techniques to more advanced adaptive online algorithms, we produce adaptive differentially private counterparts whose convergence rates depend on apriori unknown variances or parameter norms.

## [Sparse Gaussian Process Hyperparameters: Optimize or Integrate?](#)

- Vidhi Lalchand · Wessel Bruinsma · David Burt · Carl Edward Rasmussen
- abstract@[open-review](#): The kernel function and its hyperparameters are the central model selection choice in a Gaussian process (Rasmussen and Williams, 2006). Typically, the hyperparameters of the kernel are chosen by maximising the marginal likelihood, an approach known as Type-II maximum likelihood (ML-II). However, ML-II does not account for hyperparameter uncertainty, and it is well-known that this can lead to severely biased estimates and an underestimation of predictive uncertainty. While there are several works which employ fully Bayesian characterisation of GPs, relatively few propose such approaches for the sparse GPs paradigm. In this work we propose an algorithm for sparse Gaussian process regression which leverages MCMC to sample from the hyperparameter posterior within the variational inducing point framework of (Titsias, 2009). This work is closely related to (Hensman et al, 2015b) but side-steps the need to sample the inducing points, thereby significantly improving sampling efficiency in the Gaussian likelihood case. We compare this scheme against natural baselines in literature along with stochastic variational GPs (SVGPs) along with an extensive computational analysis.

## [DualCoOp: Fast Adaptation to Multi-Label Recognition with Limited Annotations](#)

- Ximeng Sun · Ping Hu · Kate Saenko
- abstract@[open-review](#): Solving multi-label recognition (MLR) for images in the low-label regime is a challenging task with many real-world applications. Recent work learns an alignment between textual and visual spaces to compensate for insufficient image labels, but loses accuracy because of the limited amount of available MLR annotations. In this work, we utilize the strong alignment of textual and visual features pretrained with millions of auxiliary image-text pairs and propose \textit{Dual Context Optimization} (DualCoOp) as a unified framework for partial-label MLR and zero-shot MLR. \ours encodes positive and negative contexts with class names as part of the linguistic input (i.e. prompts). Since \ours only introduces a very light learnable overhead upon the pretrained vision-language framework, it can quickly adapt to multi-label recognition tasks that have limited annotations and even unseen classes. Experiments on standard multi-label recognition benchmarks across two challenging low-label settings demonstrate the advantages of our approach over state-of-the-art methods. Our code will be publicly available.

## [Diversity vs. Recognizability: Human-like generalization in one-shot generative models](#)

- Victor Boutin · Lakshya Singhal · Xavier Thomas · Thomas Serre
- abstract@[open-review](#): Robust generalization to new concepts has long remained a distinctive feature of human intelligence. However, recent progress in deep generative models has now led to neural architectures capable of synthesizing novel instances of unknown visual concepts from a single training example. Yet, a more precise comparison between these models and humans is not possible because existing performance metrics for generative models (i.e., FID, IS, likelihood) are not appropriate for the one-shot generation scenario. Here, we propose a new framework to evaluate one-shot generative models along two axes: sample recognizability vs. diversity (i.e., intra-class variability). Using this framework, we perform a systematic evaluation of representative one-shot generative models on the Omniglot handwritten dataset. We first show that GAN-like and VAE-like models fall on opposite ends of the diversity-recognizability space. Extensive analyses of the effect of key model parameters further revealed that spatial attention and context integration have a linear contribution to the diversity-recognizability trade-off. In contrast, disentanglement transports the model along a parabolic curve that could be used to maximize recognizability. Using the diversity-recognizability framework, we were able to identify models and parameters that closely approximate human data.

## [Conservative Dual Policy Optimization for Efficient Model-Based Reinforcement Learning](#)

- Shenao Zhang
- abstract@[open-review](#): Provably efficient Model-Based Reinforcement Learning (MBRL) based on optimism or posterior sampling (PSRL) is ensured to attain the global optimality asymptotically by introducing the complexity measure of the model. However, the complexity might grow exponentially for the simplest nonlinear models, where global convergence is impossible within finite iterations. When the model suffers a large generalization error, which is quantitatively measured by the model complexity, the uncertainty can be large. The sampled model that current policy is greedily optimized upon will thus be unsettled, resulting in aggressive policy updates and over-exploration. In this work, we propose Conservative Dual Policy Optimization (CDPO) that involves a Referential Update and a Conservative Update. The policy is first optimized under a reference model, which imitates the mechanism of PSRL while offering more stability. A conservative range of randomness is guaranteed by maximizing the expectation of model value. Without harmful sampling procedures, CDPO can still achieve the same regret as PSRL. More importantly, CDPO enjoys monotonic policy improvement and global optimality simultaneously. Empirical results also validate the exploration efficiency of CDPO.

## [The Sample Complexity of One-Hidden-Layer Neural Networks](#)

- Gal Vardi · Ohad Shamir · Nati Srebro
- abstract@[open-review](#): We study norm-based uniform convergence bounds for neural networks, aiming at a tight understanding of how these are affected by the architecture and type of norm constraint, for the simple class of scalar-valued one-hidden-layer networks, and inputs bounded in Euclidean norm. We begin by proving that in general, controlling the spectral norm of the hidden layer weight matrix is insufficient to get uniform convergence guarantees (independent of the network width), while a stronger Frobenius norm control is sufficient, extending and improving on previous work. Motivated by the proof constructions, we identify and analyze two important settings where (perhaps surprisingly) a mere spectral norm control turns out to be sufficient: First, when the network's activation functions are sufficiently smooth (with the result extending to deeper networks); and second, for certain types of convolutional networks. In the latter setting, we study how the sample complexity is additionally affected by parameters such as the amount of overlap between patches and the overall number of patches.

## [On Margin Maximization in Linear and ReLU Networks](#)

- Gal Vardi · Ohad Shamir · Nati Srebro
- abstract@[open-review](#): The implicit bias of neural networks has been extensively studied in recent years. Lyu and Li (2019) showed that in homogeneous networks trained with the exponential or the logistic loss, gradient flow converges to a KKT point of the max margin problem in parameter space. However, that leaves open the question of whether this point will generally be an actual optimum of the max margin problem. In this paper, we study this question in detail, for several neural network architectures involving linear and ReLU activations. Perhaps surprisingly, we show that in many cases, the KKT point is not even a local optimum of the max margin problem. On the flip side, we identify multiple settings where a local or global optimum can be guaranteed.

## [Thinking Outside the Ball: Optimal Learning with Gradient Descent for Generalized Linear Stochastic Convex Optimization](#)

- Idan Amir · Roi Livni · Nati Srebro
- abstract@[open-review](#): We consider linear prediction with a convex Lipschitz loss, or more generally, stochastic convex optimization problems of generalized linear form, i.e. where each instantaneous loss is a scalar convex function of a linear function. We show that in this setting, early stopped Gradient Descent (GD), without any explicit regularization or projection, ensures excess error at most  $\$\\varepsilon\$$  (compared to the best possible with unit Euclidean norm) with an optimal, up to logarithmic factors, sample complexity of  $\$\\tilde{O}(1/\\varepsilon^2)\$$  and only  $\$\\tilde{O}(1/\\varepsilon^2)\$$  iterations. This contrasts with general stochastic convex optimization, where  $\$\\Omega(1/\\varepsilon^4)\$$  iterations are needed Amir et al. 2021. The lower iteration complexity is ensured by leveraging uniform convergence rather than stability. But instead of uniform convergence in a norm ball, which we show can guarantee suboptimal learning using  $\$\\Theta(1/\\varepsilon^4)\$$  samples, we rely on uniform convergence in a distribution-dependent ball.

## [Exponential Family Model-Based Reinforcement Learning via Score Matching](#)

- Gene Li · Junbo Li · Anmol Kabra · Nati Srebro · Zhaoran Wang · Zhioran Yang
- abstract@[open-review](#): We propose an optimistic model-based algorithm, dubbed SMRL, for finite-horizon episodic reinforcement learning (RL) when the transition model is specified by exponential family distributions with  $d$  parameters and the reward is bounded and known. SMRL uses score matching, an unnormalized density estimation technique that enables efficient estimation of the model parameter by ridge regression. Under standard regularity assumptions, SMRL achieves  $\tilde{O}(d\sqrt{H^3T})$  online regret, where  $H$  is the length of each episode and  $T$  is the total number of interactions (ignoring polynomial dependence on structural scale parameters).

## [Risk Bounds of Multi-Pass SGD for Least Squares in the Interpolation Regime](#)

- Difan Zou · Jingfeng Wu · Vladimir Braverman · Quanquan Gu · Sham Kakade
- abstract@[open-review](#): Stochastic gradient descent (SGD) has achieved great success due to its superior performance in both optimization and generalization. Most of existing generalization analyses are made for single-pass SGD, which is a less practical variant compared to the commonly-used multi-pass SGD. Besides, theoretical analyses for multi-pass SGD often concern a worst-case instance in a class of problems, which may be pessimistic to explain the superior generalization ability for some particular problem instance. The goal of this paper is to provide an instance-dependent excess risk bound of multi-pass SGD for least squares in the interpolation regime, which is expressed as a function of the iteration number, stepsize, and data covariance. We show that the excess risk of SGD can be exactly decomposed into the excess risk of GD and a positive fluctuation error, suggesting that SGD always performs worse, instance-wisely, than GD, in generalization. On the other hand, we show that although SGD needs more iterations than GD to achieve the same level of excess risk, it saves the number of stochastic gradient evaluations, and therefore is preferable in terms of computational time.

## [S4ND: Modeling Images and Videos as Multidimensional Signals with State Spaces](#)

- Eric Nguyen · Karan Goel · Albert Gu · Gordon Downs · Preety Shah · Tri Dao · Stephen Baccus · Christopher RÃ©
- abstract@[open-review](#): Visual data such as images and videos are typically modeled as discretizations of inherently continuous, multidimensional signals. Existing continuous-signal models attempt to exploit this fact by modeling the underlying signals of visual (e.g., image) data directly. However, they have not yet been able to achieve competitive performance on practical vision tasks such as large-scale image and video classification. Building on a recent line of work on deep state space models (SSMs), we propose \method, a new multidimensional SSM layer that extends SSMs' continuous-signal modeling ability to multidimensional data including images and videos. We show that S4ND can model large-scale visual data in 1D, 2D, and 3D as continuous multidimensional signals and demonstrate strong performance by simply swapping Conv2D and self-attention layers with \method layers in existing state-of-the-art models. On ImageNet-1k, \method exceeds the performance of a ViT baseline by 1.5% accuracy when training with a 1D sequence of patches, and matches ConvNeXt when modeling images in 2D. For videos, S4ND improves on an inflated 3D ConvNeXt in activity classification on HMDB-51 by 4% accuracy. S4ND implicitly learns global, continuous convolutional kernels that are resolution invariant by construction, providing an inductive bias that enables generalization across multiple resolutions. By developing a simple bandlimiting modification to S4 to overcome aliasing, S4ND achieves strong zero-shot (unseen at test time) resolution performance, e.g. achieving 88.7% accuracy on CIFAR-10 when trained on 16 \times 16 and tested on 32 \times 32 images. When trained with progressive resizing, S4ND comes within 1% of a high-resolution model while training 22% faster.

## [Tikhonov Regularization is Optimal Transport Robust under Martingale Constraints](#)

- Jiajin Li · Sirui Lin · Jose Blanchet · Viet Anh Nguyen
- abstract@[open-review](#): Distributionally robust optimization (DRO) has been shown to offer a principled way to regularize learning models. In this paper, we find that Tikhonov regularization is distributionally robust in an optimal transport sense (i.e. if an adversary chooses distributions in a suitable optimal transport neighborhood of the empirical measure), provided that suitable martingale constraints are also imposed. Further, we introduce a relaxation of the martingale constraints which not only provide a unified viewpoint to a class of existing robust methods but also lead to new regularization tools. To realize these novel tools, provably efficient computational algorithms are proposed. As a byproduct, the strong duality theorem proved in this paper can be potentially applied to other problems of independent interest.

## [Optimal Parameter-free Online Learning with Switching Cost](#)

- Zhiyu Zhang · Ashok Cutkosky · Yannis Paschalidis
- abstract@[open-review](#): Parameter-freeness in online learning refers to the adaptivity of an algorithm with respect to the optimal decision in hindsight. In this paper, we design such algorithms in the presence of switching cost - the latter penalizes the optimistic updates required by parameter-freeness, leading to a delicate design trade-off. Based on a novel dual space scaling strategy, we propose a simple yet powerful algorithm for Online Linear Optimization (OLO) with switching cost, which improves the existing suboptimal regret bound [ZCP22a] to the optimal rate. The obtained benefit is extended to the expert setting, and the practicality of our algorithm is demonstrated through a sequential investment task.

## [Doubly Robust Counterfactual Classification](#)

- Kwangho Kim · Edward Kennedy · Jose Zubizarreta
- abstract@[open-review](#): Much of the causal inference literature is concerned with estimands that can be expressed in closed form, such as the average treatment effect. In this work, we consider instead general estimands that are expressible as the solution to a nonlinear optimization problem involving counterfactuals. In particular, we study counterfactual classification as a new decision-making tool for classification under hypothetical (contrary to fact) scenarios. We propose a doubly-robust nonparametric estimator for our counterfactual classifier where we can incorporate flexible constraints. We go on to analyze rates of convergence and provide a closed-form expression for the asymptotic distribution of our estimator. Our analysis shows that the proposed estimator is robust against nuisance model misspecification, and can attain fast  $\sqrt{n}$  rates with tractable inference even when using flexible machine learning approaches. Finally, we study our methods via simulation, and apply them for recidivism risk prediction.

## [Nonlinear Sufficient Dimension Reduction with a Stochastic Neural Network](#)

- SIQI LIANG · Yan Sun · Faming Liang
- abstract@[open-review](#): Sufficient dimension reduction is a powerful tool to extract core information hidden in the high-dimensional data and has potentially many important applications in machine learning tasks. However, the existing nonlinear sufficient dimension reduction methods often lack the scalability necessary for dealing with large-scale data. We propose a new type of stochastic neural network under a rigorous probabilistic framework and show that it can be used for sufficient dimension reduction for large-scale data. The proposed stochastic neural network is trained using an adaptive stochastic gradient Markov chain Monte Carlo algorithm, whose convergence is rigorously studied in the paper as well. Through extensive experiments on real-world classification and regression problems, we show that the proposed method compares favorably with the existing state-of-the-art sufficient dimension reduction methods and is computationally more efficient for large-scale data.

## [On the Efficient Implementation of High Accuracy Optimality of Profile Maximum Likelihood](#)

- Moses Charikar Å· Zhihao Jiang Å· Kirankumar Shiragur Å· Aaron Sidford
- abstract@[open-review](#): In this paper we provide an efficient algorithm to compute a universal plug-in estimator for symmetric properties of distributions that is sample optimal up to accuracy  $\$\\epsilon \gg n^{-1/3}$ , where  $n$  is the sample size. Our estimator is based on profile-maximum-likelihood (PML) and improves upon the previous best accuracy threshold of  $\$\\epsilon \gg n^{-1/4}$  achievable by the polynomial time computable PML-based universal estimator of [\cite{ACSS20, ACSS20b}](#). Further, this reaches a theoretical limit as [\cite{Han20}](#) shows that a broad class of PML-based estimators (containing the new one we provide) are sub optimal in the regime  $\$\\epsilon \ll n^{-1/3}$ .

## [Optimal and Adaptive Monteiro-Svaiter Acceleration](#)

- Yair Carmon Å· Danielle Hausler Å· Arun Jambulapati Å· Yujia Jin Å· Aaron Sidford
- abstract@[open-review](#): We develop a variant of the Monteiro-Svaiter (MS) acceleration framework that removes the need to solve an expensive implicit equation at every iteration. Consequently, for any  $p \geq 2$  we improve the complexity of convex optimization with Lipschitz  $p$ th derivative by a logarithmic factor, matching a lower bound. We also introduce an MS subproblem solver that requires no knowledge of problem parameters, and implement it as either a second- or first-order method by solving linear systems or applying MinRes, respectively. On logistic regression problems our method outperforms previous accelerated second-order methods, but under-performs Newton's method; simply iterating our first-order adaptive subproblem solver is competitive with L-BFGS.

## [Scalable Sensitivity and Uncertainty Analyses for Causal-Effect Estimates of Continuous-Valued Interventions](#)

- Andrew Jesson Å· Alyson Douglas Å· Peter Manshausen Å· Nicolai Meinshausen Å· Philip Stier Å· Yarin Gal Å· Uri Shalit
- abstract@[open-review](#): Estimating the effects of continuous-valued interventions from observational data is a critically important task for climate science, healthcare, and economics. Recent work focuses on designing neural network architectures and regularization functions to allow for scalable estimation of average and individual-level dose-response curves from high-dimensional, large-sample data. Such methodologies assume ignorability (observation of all confounding variables) and positivity (observation of all treatment levels for every covariate value describing a set of units), assumptions problematic in the continuous treatment regime. Scalable sensitivity and uncertainty analyses to understand the ignorance induced in causal estimates when these assumptions are relaxed are less studied. Here, we develop a continuous treatment-effect marginal sensitivity model (CMSM) and derive bounds that agree with the observed data and a researcher-defined level of hidden confounding. We introduce a scalable algorithm and uncertainty-aware deep models to derive and estimate these bounds for high-dimensional, large-sample observational data. We work in concert with climate scientists interested in the climatological impacts of human emissions on cloud properties using satellite observations from the past 15 years. This problem is known to be complicated by many unobserved confounders.

## [Enhanced Bilevel Optimization via Bregman Distance](#)

- Feihu Huang Å· Junyi Li Å· Shangqian Gao Å· Heng Huang
- abstract@[open-review](#): Bilevel optimization has been recently used in many machine learning problems such as hyperparameter optimization, policy optimization, and meta learning. Although many bilevel optimization methods have been proposed, they still suffer from the high computational complexities and do not consider the more general bilevel problems with nonsmooth regularization. In the paper, thus, we propose a class of enhanced bilevel optimization methods with using Bregman distance to solve bilevel optimization problems, where the outer subproblem is nonconvex and possibly nonsmooth, and the inner subproblem is strongly convex. Specifically, we propose a bilevel optimization method based on Bregman distance (BiO-BreD) to solve deterministic bilevel problems, which achieves a lower computational complexity than the best known results. Meanwhile, we also propose a stochastic bilevel optimization method (SBiO-BreD) to solve stochastic bilevel problems based on stochastic approximated gradients and Bregman distance. Moreover, we further propose an accelerated version of SBiO-BreD method (ASBiO-BreD) using the variance-reduced technique, which can achieve a lower computational complexity than the best known computational complexity with respect to condition number  $\kappa$  and target accuracy  $\epsilon$  for finding an  $\epsilon$ -stationary point. We conduct data hyper-cleaning task and hyper-representation learning task to demonstrate that our new algorithms outperform related bilevel optimization approaches.

## [Data-Efficient Pipeline for Offline Reinforcement Learning with Limited Data](#)

- Allen Nie Å· Yannis Flet-Berliac Å· Deon Jordan Å· William Steenbergen Å· Emma Brunskill
- abstract@[open-review](#): Offline reinforcement learning (RL) can be used to improve future performance by leveraging historical data. There exist many different algorithms for offline RL, and it is well recognized that these algorithms, and their hyperparameter settings, can lead to decision policies with substantially differing performance. This prompts the need for pipelines that allow practitioners to systematically perform algorithm-hyperparameter selection for their setting. Critically, in most real-world settings, this pipeline must only involve the use of historical data. Inspired by statistical model selection methods for supervised learning, we introduce a task- and method-agnostic pipeline for automatically training, comparing, selecting, and deploying the best policy when the provided dataset is limited in size. In particular, our work highlights the importance of performing multiple data splits to produce more reliable algorithm-hyperparameter selection: while this is a common approach in supervised learning, to our knowledge, this has not been discussed in detail in the offline RL setting, and we show it can have substantial impacts when the dataset is small. Compared to alternate approaches, our proposed pipeline outputs higher-performing deployed policies from a broad range of offline policy learning algorithms and across various simulation domains in healthcare, education, and robotics. This work contributes toward the development of a general-purpose meta-algorithm for automatic algorithm-hyperparameter selection for offline RL.

## [Contact-aware Human Motion Forecasting](#)

- Wei Mao Å· miaomiao Liu Å· Richard I Hartley Å· Mathieu Salzmann
- abstract@[open-review](#): In this paper, we tackle the task of scene-aware 3D human motion forecasting, which consists of predicting future human poses given a 3D scene and a past human motion. A key challenge of this task is to ensure consistency between the human and the scene, accounting for human-scene interactions. Previous attempts to do so model such interactions only implicitly, and thus tend to produce artifacts such as ``ghost motion'' because of the lack of explicit constraints between the local poses and the global motion. Here, by contrast, we propose to explicitly model the human-scene contacts. To this end, we introduce distance-based contact maps that capture the contact relationships between every joint and every 3D scene point at each time instant. We then develop a two-stage pipeline that first predicts the future contact maps from the past ones and the scene point cloud, and then forecasts the future human poses by conditioning them on the predicted contact maps. During training, we explicitly encourage consistency between the global motion and the local poses via a prior defined using the contact maps and future poses. Our approach outperforms the state-of-the-art human motion forecasting and human synthesis methods on both synthetic and real datasets.

## [On the Limitations of Stochastic Pre-processing Defenses](#)

- Yue Gao Å· I Shumailov Å· Kassem Fawaz Å· Nicolas Papernot
- abstract@[open-review](#): Defending against adversarial examples remains an open problem. A common belief is that randomness at inference increases the cost of finding adversarial inputs. An example of such a defense is to apply a random transformation to inputs prior to feeding them to the model. In this paper, we empirically and theoretically investigate such stochastic pre-processing defenses and demonstrate that they are flawed. First, we show that most stochastic defenses are weaker than previously thought; they lack sufficient randomness to withstand even standard attacks like projected gradient descent. This casts doubt on a long-held assumption that stochastic defenses invalidate attacks designed to evade deterministic defenses and force attackers to

integrate the Expectation over Transformation (EOT) concept. Second, we show that stochastic defenses confront a trade-off between adversarial robustness and model invariance; they become less effective as the defended model acquires more invariance to their randomization. Future work will need to decouple these two effects. We also discuss implications and guidance for future research.

## [Approximate Value Equivalence](#)

- Christopher Grimm · Andre Barreto · Satinder Singh
- abstract@[open-review](#): Model-based reinforcement learning agents must make compromises about which aspects of the environment their models should capture. The value equivalence (VE) principle posits that these compromises should be made considering the model's eventual use in value-based planning. Given sets of functions and policies, a model is said to be order-\$k\$ VE to the environment if \$k\$ applications of the Bellman operators induced by the policies produce the correct result when applied to the functions. Prior work investigated the classes of models induced by VE when we vary \$k\$ and the sets of policies and functions. This gives rise to a rich collection of topological relationships and conditions under which VE models are optimal for planning. Despite this effort, relatively little is known about the planning performance of models that fail to satisfy these conditions. This is due to the rigidity of the VE formalism, as classes of VE models are defined with respect to exact constraints on their Bellman operators. This limitation gets amplified by the fact that such constraints themselves may depend on functions that can only be approximated in practice. In this work we propose an approximate theory of VE to alleviate these problems. To address these problems we propose approximate value equivalence (AVE), which extends the VE formalism by replacing equalities with error tolerances. This extension allows us to show that AVE models with respect to one set of functions are also AVE with respect to any other set of functions if we tolerate a high enough error. We can then derive bounds on the performance of VE models with respect to arbitrary sets of functions by relating them to a particular model class with known performance guarantees. In this paper we develop a theory of AVE which both extends previous topological results and enables the analysis of a wide range of VE models. We support these results by providing intuitions and discussions of their implications.

## [The price of ignorance: how much does it cost to forget noise structure in low-rank matrix estimation?](#)

- Jean Barbier · TianQi Hou · Marco Mondelli · Manuel Saenz
- abstract@[open-review](#): We consider the problem of estimating a rank-\$1\$ signal corrupted by structured rotationally invariant noise, and address the following question: \emph{how well do inference algorithms perform when the noise statistics is unknown and hence Gaussian noise is assumed?} While the matched Bayes-optimal setting with unstructured noise is well understood, the analysis of this mismatched problem is only at its premises. In this paper, we make a step towards understanding the effect of the strong source of mismatch which is the noise statistics. Our main technical contribution is the rigorous analysis of a Bayes estimator and of an approximate message passing (AMP) algorithm, both of which incorrectly assume a Gaussian setup. The first result exploits the theory of spherical integrals and of low-rank matrix perturbations; the idea behind the second one is to design and analyze an artificial AMP which, by taking advantage of the flexibility in the denoisers, is able to "correct" the mismatch. Armed with these sharp asymptotic characterizations, we unveil a rich and often unexpected phenomenology. For example, despite AMP is in principle designed to efficiently compute the Bayes estimator, the former is \emph{outperformed} by the latter in terms of mean-square error. We show that this performance gap is due to an incorrect estimation of the signal norm. In fact, when the SNR is large enough, the overlaps of the AMP and the Bayes estimator coincide, and they even match those of optimal estimators taking into account the structure of the noise.

## [Local Spatiotemporal Representation Learning for Longitudinally-consistent Neuroimage Analysis](#)

- Mengwei Ren · Neel Dey · Martin Styner · Kelly Botteron · Guido Gerig
- abstract@[open-review](#): Recent self-supervised advances in medical computer vision exploit the global and local anatomical self-similarity for pretraining prior to downstream tasks such as segmentation. However, current methods assume i.i.d. image acquisition, which is invalid in clinical study designs where follow-up longitudinal scans track subject-specific temporal changes. Further, existing self-supervised methods for medically-relevant image-to-image architectures exploit only spatial or temporal self-similarity and do so via a loss applied only at a single image-scale, with naive multi-scale spatiotemporal extensions collapsing to degenerate solutions. To these ends, this paper makes two contributions: (1) It presents a local and multi-scale spatiotemporal representation learning method for image-to-image architectures trained on longitudinal images. It exploits the spatiotemporal self-similarity of learned multi-scale intra-subject image features for pretraining and develops several feature-wise regularizations that avoid degenerate representations; (2) During finetuning, it proposes a surprisingly simple self-supervised segmentation consistency regularization to exploit intra-subject correlation. Benchmarked across various segmentation tasks, the proposed framework outperforms both well-tuned randomly-initialized baselines and current self-supervised techniques designed for both i.i.d. and longitudinal datasets. These improvements are demonstrated across both longitudinal neurodegenerative adult MRI and developing infant brain MRI and yield both higher performance and longitudinal consistency.

## [ToDD: Topological Compound Fingerprinting in Computer-Aided Drug Discovery](#)

- Anda Demir · Baris Coskunuzer · Yulia Gel · Ignacio Segovia-Dominguez · Yuzhou Chen · Bulent Kiziltan
- abstract@[open-review](#): In computer-aided drug discovery (CADD), virtual screening (VS) is used for comparing a library of compounds against known active ligands to identify the drug candidates that are most likely to bind to a molecular target. Most VS methods to date have focused on using canonical compound representations (e.g., SMILES strings, Morgan fingerprints) or generating alternative fingerprints of the compounds by training progressively more complex variational autoencoders (VAEs) and graph neural networks (GNNs). Although VAEs and GNNs led to significant improvements in VS performance, these methods suffer from reduced performance when scaling to large virtual compound datasets. The performance of these methods has shown only incremental improvements in the past few years. To address this problem, we developed a novel method using multiparameter persistence (MP) homology that produces topological fingerprints of the compounds as multidimensional vectors. Our primary contribution is framing the VS process as a new topology-based graph ranking problem by partitioning a compound into chemical substructures informed by the periodic properties of its atoms and extracting their persistent homology features at multiple resolution levels. We show that the margin loss fine-tuning of pretrained Triplet networks attains highly competitive results in differentiating between compounds in the embedding space and ranking their likelihood of becoming effective drug candidates. We further establish theoretical guarantees for the stability properties of our proposed MP signatures, and demonstrate that our models, enhanced by the MP signatures, outperform state-of-the-art methods on benchmark datasets by a wide and highly statistically significant margin (e.g., 93% gain for Cleves-Jain and 54% gain for DUD-E Diverse dataset).

## [Score-Based Diffusion meets Annealed Importance Sampling](#)

- Arnaud Doucet · Will Grathwohl · Alexander Matthews · Heiko Strathmann
- abstract@[open-review](#): More than twenty years after its introduction, Annealed Importance Sampling (AIS) remains one of the most effective methods for marginal likelihood estimation. It relies on a sequence of distributions interpolating between a tractable initial distribution and the target distribution of interest which we simulate from approximately using a non-homogeneous Markov chain. To obtain an importance sampling estimate of the marginal likelihood, AIS introduces an extended target distribution to reweight the Markov chain proposal. While much effort has been devoted to improving the proposal distribution used by AIS, by changing the intermediate distributions and corresponding Markov kernels, an underappreciated issue is that AIS uses a convenient but suboptimal extended target distribution. This can hinder its performance. We here leverage recent progress in score-based generative modeling (SGM) to approximate the optimal extended target distribution for AIS proposals corresponding to the discretization of Langevin and Hamiltonian dynamics. We demonstrate these novel, differentiable, AIS procedures on a number of synthetic benchmark distributions and variational auto-encoders.

## [HorNet: Efficient High-Order Spatial Interactions with Recursive Gated Convolutions](#)

- Yongming Rao · Wenliang Zhao · Yansong Tang · Jie Zhou · Ser Nam Lim · Jiwen Lu
- abstract@[open-review](#): Recent progress in vision Transformers exhibits great success in various tasks driven by the new spatial modeling mechanism based on dot-product self-attention. In this paper, we show that the key ingredients behind the vision Transformers, namely input-adaptive, long-range and high-order spatial interactions, can also be efficiently implemented with a convolution-based framework. We present the Recursive Gated Convolution ( $\text{Conv}$ ) that performs high-order spatial interactions with gated convolutions and recursive designs. The new operation is highly flexible and customizable, which is compatible with various variants of convolution and extends the two-order interactions in self-attention to arbitrary orders without introducing significant extra computation.  $\text{Conv}$  can serve as a plug-and-play module to improve various vision Transformers and convolution-based models. Based on the operation, we construct a new family of generic vision backbones named HorNet. Extensive experiments on ImageNet classification, COCO object detection and ADE20K semantic segmentation show HorNet outperform Swin Transformers and ConvNeXt by a significant margin with similar overall architecture and training configurations. HorNet also shows favorable scalability to more training data and a larger model size. Apart from the effectiveness in visual encoders, we also show  $\text{Conv}$  can be applied to task-specific decoders and consistently improve dense prediction performance with less computation. Our results demonstrate that  $\text{Conv}$  can be a new basic module for visual modeling that effectively combines the merits of both vision Transformers and CNNs.

## [M\\$^4\\$I: Multi-modal Models Membership Inference](#)

- Pingyi Hu · Zihan Wang · Ruoxi Sun · Hu Wang · Minhui Xue
- abstract@[open-review](#): With the development of machine learning techniques, the attention of research has been moved from single-modal learning to multi-modal learning, as real-world data exist in the form of different modalities. However, multi-modal models often carry more information than single-modal models and they are usually applied in sensitive scenarios, such as medical report generation or disease identification. Compared with the existing membership inference against machine learning classifiers, we focus on the problem that the input and output of the multi-modal models are in different modalities, such as image captioning. This work studies the privacy leakage of multi-modal models through the lens of membership inference attack, a process of determining whether a data record involves in the model training process or not. To achieve this, we propose Multi-modal Models Membership Inference (M\$^4\$I) with two attack methods to infer the membership status, named metric-based (MB) M\$^4\$I and feature-based (FB) M\$^4\$I, respectively. More specifically, MB M\$^4\$I adopts similarity metrics while attacking to infer target data membership. FB M\$^4\$I uses a pre-trained shadow multi-modal feature extractor to achieve the purpose of data inference attack by comparing the similarities from extracted input and output features. Extensive experimental results show that both attack methods can achieve strong performances. Respectively, 72.5% and 94.83% of attack success rates on average can be obtained under unrestricted scenarios. Moreover, we evaluate multiple defense mechanisms against our attacks. The source code of M\$^4\$I attacks is publicly available at <https://github.com/MultimodalMI/Multimodal-membership-inference.git>.

## [Green Hierarchical Vision Transformer for Masked Image Modeling](#)

- Lang Huang · Shan You · Mingkai Zheng · Fei Wang · Chen Qian · Toshihiko Yamasaki
- abstract@[open-review](#): We present an efficient approach for Masked Image Modeling (MIM) with hierarchical Vision Transformers (ViTs), e.g., Swin Transformer, allowing the hierarchical ViTs to discard masked patches and operate only on the visible ones. Our approach consists of two key components. First, for the window attention, we design a Group Window Attention scheme following the Divide-and-Conquer strategy. To mitigate the quadratic complexity of the self-attention w.r.t. the number of patches, group attention encourages a uniform partition that visible patches within each local window of arbitrary size can be grouped with equal size, where masked self-attention is then performed within each group. Second, we further improve the grouping strategy via the Dynamic Programming algorithm to minimize the overall computation cost of the attention on the grouped patches. As a result, MIM now can work on hierarchical ViTs in a green and efficient way. For example, we can train the hierarchical ViTs about 2x faster and reduce the GPU memory usage by 60%, while still enjoying competitive performance on ImageNet classification and the superiority on downstream COCO object detection benchmarks.

## [Deep Generalized Schrödinger Bridge](#)

- Guan-Horng Liu · Tianrong Chen · Oswin So · Evangelos Theodorou
- abstract@[open-review](#): Mean-Field Game (MFG) serves as a crucial mathematical framework in modeling the collective behavior of individual agents interacting stochastically with a large population. In this work, we aim at solving a challenging class of MFGs in which the differentiability of these interacting preferences needs not available to the solver, and the population is urged to converge exactly to some desired distribution. These setups are, despite being well-motivated for practical purposes, complicated enough to paralyze most (deep) numerical solvers. Nevertheless, we show that Schrödinger Bridge “” as an entropy-regularized optimal transport model “” can be generalized to accepting mean-field structures, hence solving these MFGs. This is achieved via the application of Forward-Backward Stochastic Differential Equations theory, which, intriguingly, leads to a computational framework with a similar structure to Temporal Difference learning. As such, it opens up novel algorithmic connections to Deep Reinforcement Learning that we leverage for facilitating practical training. Our method, named Deep Generalized Schrödinger Bridge (DeepGSB), not only outperforms prior methods in solving classical population navigation MFGs, but is also capable of solving 1000-dimensional opinion depolarization, setting a new state-of-the-art numerical solver for high-dimensional MFGs.

## [One-shot Neural Backdoor Erasing via Adversarial Weight Masking](#)

- Shuwen Chai · Jinghui Chen
- abstract@[open-review](#): Recent studies show that despite achieving high accuracy on a number of real-world applications, deep neural networks (DNNs) can be backdoored: by injecting triggered data samples into the training dataset, the adversary can mislead the trained model into classifying any test data to the target class as long as the trigger pattern is presented. To nullify such backdoor threats, various methods have been proposed. Particularly, a line of research aims to purify the potentially compromised model. However, one major limitation of this line of work is the requirement to access sufficient original training data: the purifying performance is a lot worse when the available training data is limited. In this work, we propose Adversarial Weight Masking (AWM), a novel method capable of erasing the neural backdoors even in the one-shot setting. The key idea behind our method is to formulate this into a min-max optimization problem: first, adversarially recover the non-robust perturbation patterns and then (soft) mask the network weights that are sensitive to the recovered patterns. Comprehensive evaluations of several benchmark datasets suggest that AWM can largely improve the purifying effects over other state-of-the-art methods on various available training dataset sizes.

## [A Near-Optimal Primal-Dual Method for Off-Policy Learning in CMDP](#)

- Fan Chen · Junyu Zhang · Zaiwen Wen
- abstract@[open-review](#): As an important framework for safe Reinforcement Learning, the Constrained Markov Decision Process (CMDP) has been extensively studied in the recent literature. However, despite the rich results under various on-policy learning settings, there still lacks some essential understanding of the offline CMDP problems, in terms of both the algorithm design and the information theoretic sample complexity lower bound. In this paper, we focus on solving the CMDP problems where only offline data are available. By adopting the concept of the single-policy concentrability coefficient  $C$ , we establish an  $\Omega(\min\{\frac{1}{\gamma}, \frac{1}{C}\} + \frac{1}{C})$  sample complexity lower bound for the offline CMDP problem, where  $I$  stands for the number of constraints. By introducing a simple but novel deviation control mechanism, we propose a near-optimal primal-dual learning algorithm called DPDL. This algorithm provably guarantees zero constraint

violation and its sample complexity matches the above lower bound except for an  $\tilde{O}((1-\gamma)^{-1})$  factor. Comprehensive discussion on how to deal with the unknown constant  $C^*$  and the potential asynchronous structure on the offline dataset are also included.

## [OST: Improving Generalization of DeepFake Detection via One-Shot Test-Time Training](#)

- Liang Chen · Yong Zhang · Yibing Song · Jue Wang · Lingqiao Liu
- abstract@[open-review](#): State-of-the-art deepfake detectors perform well in identifying forgeries when they are evaluated on a test set similar to the training set, but struggle to maintain good performance when the test forgeries exhibit different characteristics from the training images e.g., forgeries are created by unseen deepfake methods. Such a weak generalization capability hinders the applicability of deepfake detectors. In this paper, we introduce a new learning paradigm specially designed for the generalizable deepfake detection task. Our key idea is to construct a test-sample-specific auxiliary task to update the model before applying it to the sample. Specifically, we synthesize pseudo-training samples from each test image and create a test-time training objective to update the model. Moreover, we proposed to leverage meta-learning to ensure that a fast single-step test-time gradient descent, dubbed one-shot test-time training (OST), can be sufficient for good deepfake detection performance. Extensive results across several benchmark datasets demonstrate that our approach performs favorably against existing arts in terms of generalization to unseen data and robustness to different post-processing steps.

## [Deep Hierarchical Planning from Pixels](#)

- Danijar Hafner · Kuang-Huei Lee · Ian Fischer · Pieter Abbeel
- abstract@[open-review](#): Intelligent agents need to select long sequences of actions to solve complex tasks. While humans easily break down tasks into subgoals and reach them through millions of muscle commands, current artificial intelligence is limited to tasks with horizons of a few hundred decisions, despite large compute budgets. Research on hierarchical reinforcement learning aims to overcome this limitation but has proven to be challenging, current methods rely on manually specified goal spaces or subtasks, and no general solution exists. We introduce Director, a practical method for learning hierarchical behaviors directly from pixels by planning inside the latent space of a learned world model. The high-level policy maximizes task and exploration rewards by selecting latent goals and the low-level policy learns to achieve the goals. Despite operating in latent space, the decisions are interpretable because the world model can decode goals into images for visualization. Director learns successful behaviors across a wide range of environments, including visual control, Atari games, and DMLab levels and outperforms exploration methods on tasks with very sparse rewards, including 3D maze traversal with a quadruped robot from an egocentric camera and proprioception, without access to the global position or top-down view used by prior work.

## [CASA: Category-agnostic Skeletal Animal Reconstruction](#)

- Yuefan Wu · Zeyuan Chen · Shaowei Liu · Zhongzheng Ren · Shenlong Wang
- abstract@[open-review](#): Recovering a skeletal shape from a monocular video is a longstanding challenge. Prevailing nonrigid animal reconstruction methods often adopt a control-point driven animation model and optimize bone transforms individually without considering skeletal topology, yielding unsatisfactory shape and articulation. In contrast, humans can easily infer the articulation structure of an unknown character by associating it with a seen articulated object in their memory. Inspired by this fact, we present CASA, a novel category-agnostic articulated animal reconstruction method. Our method consists of two components, a video-to-shape retrieval process and a neural inverse graphics framework. During inference, CASA first finds a matched articulated shape from a 3D character assets bank so that the input video scores highly with the rendered image, according to a pretrained image-language model. It then integrates the retrieved character into an inverse graphics framework and jointly infers the shape deformation, skeleton structure, and skinning weights through optimization. Experiments validate the efficacy of our method in shape reconstruction and articulation. We further show that we can use the resulting skeletal-animated character for re-animation.

## [When does SGD favor flat minima? A quantitative characterization via linear stability](#)

- Lei Wu · Mingze Wang · Weijie Su
- abstract@[open-review](#): The observation that stochastic gradient descent (SGD) tends to select flat minima has played a fundamental role in understanding implicit regularization of SGD and guiding the tuning of hyperparameters. In this paper, we provide a quantitative explanation of this phenomenon by relating the particular noise structure of SGD to its *linear stability* (Wu et al., 2018). Specifically, we consider training over-parameterized models with square loss. We prove that if a global minimum  $\theta^*$  is *linearly stable* for SGD, then it must satisfy  $\|\nabla H(\theta^*)\|_F \leq O(\sqrt{B}\eta)$ , where  $\|\nabla H(\theta^*)\|_F$ ,  $B$ ,  $\eta$  denote the Frobenius norm of Hessian at  $\theta^*$ , batch size, and learning rate, respectively. Otherwise, SGD will escape from that minimum exponentially fast. Hence, for minima accessible to SGD, the flatness---as measured by the Frobenius norm of the Hessian---is bounded independently of the model size and sample size. The key to obtaining these results is exploiting the particular geometry awareness of SGD noise: 1) the noise magnitude is proportional to loss value; 2) the noise directions concentrate in the sharp directions of local landscape. This property of SGD noise provably holds for linear networks and random feature models (RFMs) and is empirically verified for nonlinear networks. Moreover, the validity and practical relevance of our theoretical findings are justified by extensive numerical experiments.

## [Muffliato: Peer-to-Peer Privacy Amplification for Decentralized Optimization and Averaging](#)

- Edwige Cyffers · Mathieu Even · Aurélien Bellet · Laurent Massoulié
- abstract@[open-review](#): Decentralized optimization is increasingly popular in machine learning for its scalability and efficiency. Intuitively, it should also provide better privacy guarantees, as nodes only observe the messages sent by their neighbors in the network graph. But formalizing and quantifying this gain is challenging: existing results are typically limited to Local Differential Privacy (LDP) guarantees that overlook the advantages of decentralization. In this work, we introduce pairwise network differential privacy, a relaxation of LDP that captures the fact that the privacy leakage from a node  $u$  to a node  $v$  may depend on their relative position in the graph. We then analyze the combination of local noise injection with (simple or randomized) gossip averaging protocols on fixed and random communication graphs. We also derive a differentially private decentralized optimization algorithm that alternates between local gradient descent steps and gossip averaging. Our results show that our algorithms amplify privacy guarantees as a function of the distance between nodes in the graph, matching the privacy-utility trade-off of the trusted curator, up to factors that explicitly depend on the graph topology. Finally, we illustrate our privacy gains with experiments on synthetic and real-world datasets.

## [Estimating and Explaining Model Performance When Both Covariates and Labels Shift](#)

- Lingjiao Chen · Matei Zaharia · James Zou
- abstract@[open-review](#): Deployed machine learning (ML) models often encounter new user data that differs from their training data. Therefore, estimating how well a given model might perform on the new data is an important step toward reliable ML applications. This is very challenging, however, as the data distribution can change in flexible ways, and we may not have any labels on the new data, which is often the case in monitoring settings. In this paper, we propose a new distribution shift model, Sparse Joint Shift (SJS), which considers the joint shift of both labels and a few features. This unifies and generalizes several existing shift models including label shift and sparse covariate shift, where only marginal feature or label distribution shifts are considered. We describe mathematical conditions under which SJS is identifiable. We further propose SEES, an algorithmic framework to characterize the distribution shift under SJS and to estimate a model's performance on new data without any labels. We conduct extensive experiments on several real-world datasets with various ML models. Across different datasets and distribution shifts, SEES achieves significant (up to an order of magnitude) shift estimation error improvements over existing approaches.

## [Adversarial Training with Complementary Labels: On the Benefit of Gradually Informative Attacks](#)

- Jianan Zhou · Jianing Zhu · Jingfeng ZHANG · Tongliang Liu · Gang Niu · Bo Han · Masashi Sugiyama
- abstract@[open-review](#): Adversarial training (AT) with imperfect supervision is significant but receives limited attention. To push AT towards more practical scenarios, we explore a new setting, i.e., AT with complementary labels (CLs), which specify a class that a data sample does not belong to. However, the direct combination of AT with existing methods for CLs results in consistent failure, but not on a simple baseline of two-stage training which consists of a complementary learning phase and an AT phase separately. In this paper, we further explore the phenomenon and identify the underlying challenges of AT with CLs as intractable adversarial optimization and low-quality adversarial examples. To address the above problems, we propose a new learning strategy using gradually informative attacks, which consists of two critical components: 1) Warm-up Attack (Warm-up) gently raises the adversarial perturbation budgets to ease the adversarial optimization with CLs; 2) Pseudo-Label Attack (PLA) incorporates the progressively informative model predictions into a corrected complementary loss. Extensive experiments are conducted to demonstrate the effectiveness of our method on a range of benchmarked datasets.

## [Text-driven Photorealistic 3D Stylization For Arbitrary Meshes](#)

- yongwei chen · chen rui · Jiabao Lei · Yabin Zhang · Kui Jia
- abstract@[open-review](#): Creation of 3D content by stylization is a promising yet challenging problem in computer vision and graphics research. In this work, we focus on stylizing photorealistic appearance renderings of a given surface mesh of arbitrary topology. Motivated by the recent surge of cross-modal supervision of the Contrastive Language-Image Pre-training (CLIP) model, we propose to transfer the appearance style of a given 3D shape according to a text prompt in a photorealistic manner. Technically, we propose to disentangle the appearance style as the spatially varying bidirectional reflectance distribution function, the local geometric variation, and the lighting condition, which are jointly optimized, via supervision of the CLIP loss, by a spherical Gaussians based differentiable renderer. As such, our method enables photorealistic 3D style transfer by automatically predicting reflectance effects even for bare, low-quality meshes, without training on a task-specific dataset. Extensive experiments show that our method outperforms existing methods of text-driven 3D style transfer in terms of photorealistic quality, consistency of 3D geometry, and robustness when stylizing low-quality meshes.

## [ZARTS: On Zero-order Optimization for Neural Architecture Search](#)

- Xiaoxing Wang · Wenzuan Guo · Jianlin Su · Xiaokang Yang · Junchi Yan
- abstract@[open-review](#): Differentiable architecture search (DARTS) has been a popular one-shot paradigm for NAS due to its high efficiency. It introduces trainable architecture parameters to represent the importance of candidate operations and proposes first/second-order approximation to estimate their gradients, making it possible to solve NAS by gradient descent algorithm. However, our in-depth empirical results show that the approximation often distorts the loss landscape, leading to the biased objective to optimize and, in turn, inaccurate gradient estimation for architecture parameters. This work turns to zero-order optimization and proposes a novel NAS scheme, called ZARTS, to search without enforcing the above approximation. Specifically, three representative zero-order optimization methods are introduced: RS, MGS, and GLD, among which MGS performs best by balancing the accuracy and speed. Moreover, we explore the connections between RS/MGS and gradient descent algorithm and show that our ZARTS can be seen as a robust gradient-free counterpart to DARTS. Extensive experiments on multiple datasets and search spaces show the remarkable performance of our method. In particular, results on 12 benchmarks verify the outstanding robustness of ZARTS, where the performance of DARTS collapses due to its known instability issue. Also, we search on the search space of DARTS to compare with peer methods, and our discovered architecture achieves 97.54% accuracy on CIFAR-10 and 75.7% top-1 accuracy on ImageNet. Finally, we combine our ZARTS with three orthogonal variants of DARTS for faster search speed and better performance.

## [Giga-scale Kernel Matrix-Vector Multiplication on GPU](#)

- Robert Hu · Siu Lun Chau · Dino Sejdinovic · Joan Glaunès
- abstract@[open-review](#): Kernel matrix-vector multiplication (KMVM) is a foundational operation in machine learning and scientific computing. However, as KMVM tends to scale quadratically in both memory and time, applications are often limited by these computational constraints. In this paper, we propose a novel approximation procedure coined \textit{Faster-Fast and Free Memory Method} (\textit{FFF}) to address these scaling issues of KMVM for tall- $\sim(10^8 \text{sim } 10^9)$  and skinny- $\sim(D \leq 7)$  data. Extensive experiments demonstrate that \textit{FFF} has empirical \textit{linear time and memory} complexity with a relative error of order  $10^{-3}$  and can compute a full KMVM for a billion points \textit{in under a minute} on a high-end GPU, leading to a significant speed-up in comparison to existing CPU methods. We demonstrate the utility of our procedure by applying it as a drop-in for the state-of-the-art GPU-based linear solver FALKON, \textit{improving speed 1.5-5.5 times} at the cost of  $<1\%$  drop in accuracy. We further demonstrate competitive results on \textit{Gaussian Process regression} coupled with significant speedups on a variety of real-world datasets.

## [Self-Supervised Learning via Maximum Entropy Coding](#)

- Xin Liu · Zhongdao Wang · Ya-Li Li · Shengjin Wang
- abstract@[open-review](#): A mainstream type of current self-supervised learning methods pursues a general-purpose representation that can be well transferred to downstream tasks, typically by optimizing on a given pretext task such as instance discrimination. In this work, we argue that existing pretext tasks inevitably introduce biases into the learned representation, which in turn leads to biased transfer performance on various downstream tasks. To cope with this issue, we propose Maximum Entropy Coding (MEC), a more principled objective that explicitly optimizes on the structure of the representation, so that the learned representation is less biased and thus generalizes better to unseen downstream tasks. Inspired by the principle of maximum entropy in information theory, we hypothesize that a generalizable representation should be the one that admits the maximum entropy among all plausible representations. To make the objective end-to-end trainable, we propose to leverage the minimal coding length in lossy data coding as a computationally tractable surrogate for the entropy, and further derive a scalable reformulation of the objective that allows fast computation. Extensive experiments demonstrate that MEC learns a more generalizable representation than previous methods based on specific pretext tasks. It achieves state-of-the-art performance consistently on various downstream tasks, including not only ImageNet linear probe, but also semi-supervised classification, object detection, instance segmentation, and object tracking. Interestingly, we show that existing batch-wise (e.g., SimSiam) and feature-wise (e.g., Barlow Twins) objectives could be seen equivalent to low-order approximations of MEC. Code will be released.

## [SHINE: SubHypergraph Inductive Neural nEtworK](#)

- Yuan Luo
- abstract@[open-review](#): Hypergraph neural networks can model multi-way connections that are beyond pairwise associations among nodes of the graphs. Multi-way connections are common in many real-world applications and, in particular, genetic medicine. In particular, genetic pathways or broadly speaking gene sets encode relationships among multiple genes that collectively drive a molecular function, which can be naturally modeled as hyperedges connecting all involved nodes (e.g., genes). Thus, hypergraph-guided embedding can capture functional relations in learned representations. Existing hypergraph neural network models often focus on node-level or graph-level inference. There is an unmet need in learning powerful representations of subgraphs of hypergraphs in real-world applications. For example, a cancer patient can be viewed as a subgraph of genes harboring mutations in the patient, while all the genes are connected by hyperedges that correspond to pathways representing specific molecular functions. To achieve accurate inductive subgraph prediction, we propose SubHypergraph Inductive Neural nEtworK (SHINE). SHINE uses informative genetic pathways that encode molecular functions as hyperedges to connect genes as nodes. SHINE jointly optimizes the objectives of end-to-end subgraph classification and hypergraph nodes' similarity regularization. SHINE simultaneously learns representations for both genes and pathways using strongly dual attention

message passing. These learned representations are then aggregated via a subgraph attention layer to derive a subgraph representation, which is used in training a multilayer perceptron for subgraph inferencing. We evaluated SHINE against a wide array of state-of-the-art hypergraph neural networks, XGBoost, NMF and polygenic risk score models, using diverse large scale NGS and curated datasets. SHINE outperformed all comparison models significantly, and yielded interpretable models with functional insights on molecular mechanisms of diseases.

## [Learning Optical Flow From Continuous Spike Streams](#)

- Rui Zhao · Ruiqin Xiong · Jing Zhao · Zhaofei Yu · Xiaopeng Fan · Tiejun Huang
- abstract@[open-review](#): Spiking camera is an emerging bio-inspired vision sensor with ultra-high temporal resolution. It records scenes by accumulating photons and outputting continuous binary spike streams. Optical flow is a key task for spiking cameras and their applications. A previous attempt has been made for spike-based optical flow. However, the previous work only focuses on motion between two moments, and it uses graphics-based data for training, whose generalization is limited. In this paper, we propose a tailored network, Spike2Flow that extracts information from binary spikes with temporal-spatial representation based on the differential of spike firing time and spatial information aggregation. The network utilizes continuous motion clues through joint correlation decoding. Besides, a new dataset with real-world scenes is proposed for better generalization. Experimental results show that our approach achieves state-of-the-art performance on existing synthetic datasets and real data captured by spiking cameras. The source code and dataset will be publicly available after publication.

## [Unsupervised Learning of Algebraic Structure from Stationary Time Sequences](#)

- Takeru Miyato · Masanori Koyama · Kenji Fukumizu
- abstract@[open-review](#): Recently, there is a surge of interest in the data-driven learning of underlying simple relations and symmetry in the dataset. However, finding such a symmetry often requires structural assumptions about either a dataset or a model, or both. In this study, we investigate the symmetry that can be learned from time sequences of length at least three. We show that, if each time sequence in the dataset has a certain stationary property (e.g. constant velocity, constant acceleration), the algebraic structure that helps us extrapolate far into the future can be discovered without supervision. We will showcase our method from both empirical and theoretical perspectives. We will also show that, when the representation is trained so that the future observation can be predicted well by a linear transition in the latent space, the hidden disentangled structure of the dataset naturally emerges as a by-product through simultaneous block-diagonalization. Our result suggests that finding a simple structured relation and learning a model with extrapolation capability are two sides of the same coin.

## [A Conditional Randomization Test for Sparse Logistic Regression in High-Dimension](#)

- Binh T. Nguyen · Bertrand Thirion · Sylvain Arlot
- abstract@[open-review](#): Identifying the relevant variables for a classification model with correct confidence levels is a central but difficult task in high-dimension. Despite the core role of sparse logistic regression in statistics and machine learning, it still lacks a good solution for accurate inference in the regime where the number of features  $p$  is as large as or larger than the number of samples  $n$ . Here we tackle this problem by improving the Conditional Randomization Test (CRT). The original CRT algorithm shows promise as a way to output p-values while making few assumptions on the distribution of the test statistics. As it comes with a prohibitive computational cost even in mildly high-dimensional problems, faster solutions based on distillation have been proposed. Yet, they rely on unrealistic hypotheses and result in low-power solutions. To improve this, we propose `CRT-logit`, an algorithm that combines a variable-distillation step and a decorrelation step that takes into account the geometry of  $\ell_1$ -penalized logistic regression problem. We provide a theoretical analysis of this procedure, and demonstrate its effectiveness on simulations, along with experiments on large-scale brain-imaging and genomics datasets.

## [SegViT: Semantic Segmentation with Plain Vision Transformers](#)

- Bowen Zhang · Zhi Tian · Quan Tang · Xiangxiang Chu · Xiaolin Wei · Chunhua Shen · Yifan Liu
- abstract@[open-review](#): We explore the capability of plain Vision Transformers (ViTs) for semantic segmentation and propose the SegViT. Previous ViT-based segmentation networks usually learn a pixel-level representation from the output of the ViT. Differently, we make use of the fundamental component – “attention mechanism, to generate masks for semantic segmentation. Specifically, we propose the Attention-to-Mask (ATM) module, in which the similarity maps between a set of learnable class tokens and the spatial feature maps are transferred to the segmentation masks. Experiments show that our proposed SegViT using the ATM module outperforms its counterparts using the plain ViT backbone on the ADE20K dataset and achieves new state-of-the-art performance on COCO-Stuff-10K and PASCAL-Context datasets. Furthermore, to reduce the computational cost of the ViT backbone, we propose query-based down-sampling (QD) and query-based up-sampling (QU) to build a Shrunk structure. With our Shrunk structure, the model can save up to 40% computations while maintaining competitive performance.

## [VITA: Video Instance Segmentation via Object Token Association](#)

- Miran Heo · Sukjun Hwang · Seoung Wug Oh · Joon-Young Lee · Seon Joo Kim
- abstract@[open-review](#): We introduce a novel paradigm for offline Video Instance Segmentation (VIS), based on the hypothesis that explicit object-oriented information can be a strong clue for understanding the context of the entire sequence. To this end, we propose VITA, a simple structure built on top of an off-the-shelf Transformer-based image instance segmentation model. Specifically, we use an image detector as a means of distilling object-specific contexts into object tokens. VITA accomplishes video-level understanding by associating frame-level object tokens without using spatio-temporal backbone features. By effectively building relationships between objects using the condensed information, VITA achieves the state-of-the-art on VIS benchmarks with ResNet-50 backbone: 49.8 AP, 45.7 AP on YouTube-VIS 2019 & 2021 and 19.6 AP on OVIS. Moreover, thanks to its object token-based structure that is disjoint from the backbone features, VITA shows several practical advantages that previous offline VIS methods have not explored - handling long and high-resolution videos with a common GPU and freezing a frame-level detector trained on image domain. Code will be made available.

## [SAVi++: Towards End-to-End Object-Centric Learning from Real-World Videos](#)

- Gamaleldin Elsayed · Aravindh Mahendran · Sjoerd van Steenkiste · Klaus Greff · Michael Mozer · Thomas Kipf
- abstract@[open-review](#): The visual world can be parsimoniously characterized in terms of distinct entities with sparse interactions. Discovering this compositional structure in dynamic visual scenes has proven challenging for end-to-end computer vision approaches unless explicit instance-level supervision is provided. Slot-based models leveraging motion cues have recently shown great promise in learning to represent, segment, and track objects without direct supervision, but they still fail to scale to complex real-world multi-object videos. In an effort to bridge this gap, we take inspiration from human development and hypothesize that information about scene geometry in the form of depth signals can facilitate object-centric learning. We introduce SAVi++, an object-centric video model which is trained to predict depth signals from a slot-based video representation. By further leveraging best practices for model scaling, we are able to train SAVi++ to segment complex dynamic scenes recorded with moving cameras, containing both static and moving objects of diverse appearance on naturalistic backgrounds, without the need for segmentation supervision. Finally, we demonstrate that by using sparse depth signals obtained from LiDAR, SAVi++ is able to learn emergent object segmentation and tracking from videos in the real-world Waymo Open dataset.

## [Resource-Adaptive Federated Learning with All-In-One Neural Composition](#)

- Yiqun Mei · Pengfei Guo · Mo Zhou · Vishal Patel
- abstract@[open-review](#): Conventional Federated Learning (FL) systems inherently assume a uniform processing capacity among clients for deployed models. However, diverse client hardware often leads to varying computation resources in practice. Such system heterogeneity results in an inevitable trade-off between model complexity and data accessibility as a bottleneck. To avoid such a dilemma and achieve resource-adaptive federated learning, we introduce a simple yet effective mechanism, termed All-In-One Neural Composition, to systematically support training complexity-adjustable models with flexible resource adaption. It is able to efficiently construct models at various complexities using one unified neural basis shared among clients, instead of pruning the global model into local ones. The proposed mechanism endows the system with unhindered access to the full range of knowledge scattered across clients and generalizes existing pruning-based solutions by allowing soft and learnable extraction of low footprint models. Extensive experiment results on popular FL benchmarks demonstrate the effectiveness of our approach. The resulting FL system empowered by our All-In-One Neural Composition, called FLANC, manifests consistent performance gains across diverse system/data heterogeneous setups while keeping high efficiency in computation and communication.

## [Behavior Transformers: Cloning \\$k\\$ modes with one stone](#)

- Nur Muhammad Shafullah · Zichen Cui · Ariuntuya Altanzaya · Lerrel Pinto
- abstract@[open-review](#): While behavior learning has made impressive progress in recent times, it lags behind computer vision and natural language processing due to its inability to leverage large, human-generated datasets. Human behavior has a wide variance, multiple modes, and human demonstrations naturally do not come with reward labels. These properties limit the applicability of current methods in Offline RL and Behavioral Cloning to learn from large, pre-collected datasets. In this work, we present Behavior Transformer (BeT), a new technique to model unlabeled demonstration data with multiple modes. BeT retrofits standard transformer architectures with action discretization coupled with a multi-task action correction inspired by offset prediction in object detection. This allows us to leverage the multi-modal modeling ability of modern transformers to predict multi-modal continuous actions. We experimentally evaluate BeT on a variety of robotic manipulation and self-driving behavior datasets. We show that BeT significantly improves over prior state-of-the-art work on solving demonstrated tasks while capturing the major modes present in the pre-collected datasets. Finally, through an extensive ablation study, we further analyze the importance of every crucial component in BeT. Videos of behavior generated by BeT are available here: <https://submission0.github.io>.

## [Near-Optimal No-Regret Learning Dynamics for General Convex Games](#)

- Gabriele Farina · Ioannis Anagnostides · Haipeng Luo · Chung-Wei Lee · Christian Kroer · Tuomas Sandholm
- abstract@[open-review](#): A recent line of work has established uncoupled learning dynamics such that, when employed by all players in a game, each player's regret after  $T$  repetitions grows polylogarithmically in  $T$ , an exponential improvement over the traditional guarantees within the no-regret framework. However, so far these results have only been limited to certain classes of games with structured strategy spaces---such as normal-form and extensive-form games. The question as to whether  $\mathcal{O}(\text{polylog } T)$  regret bounds can be obtained for general convex and compact strategy sets---as is the case in many fundamental models in economics and multiagent systems---while retaining efficient strategy updates is an important question. In this paper, we answer this in the positive by establishing the first uncoupled learning algorithm with  $\mathcal{O}(\log T)$  per-player regret in general convex games, that is, games with concave utility functions supported on arbitrary convex and compact strategy sets. Our learning dynamics are based on an instantiation of optimistic follow-the-regularized-leader over an appropriately lifted space using a self-concordant regularizer that is peculiarly not a barrier for the feasible region. Our learning dynamics are efficiently implementable given access to a proximal oracle for the convex strategy set, leading to  $\mathcal{O}(\log \log T)$  per-iteration complexity; we also give extensions when access to only a linear optimization oracle is assumed. Finally, we adapt our dynamics to guarantee  $\mathcal{O}(\sqrt{T})$  regret in the adversarial regime. Even in those special cases where prior results apply, our algorithm improves over the state-of-the-art regret bounds either in terms of the dependence on the number of iterations or on the dimension of the strategy sets.

## [Single-pass Streaming Lower Bounds for Multi-armed Bandits Exploration with Instance-sensitive Sample Complexity](#)

- Sepehr Assadi · Chen Wang
- abstract@[open-review](#): Motivated by applications to process massive datasets, we study streaming algorithms for pure exploration in Stochastic Multi Armed Bandits (MABs). This problem was first formulated by Assadi and Wang [STOC 2020] as follows: A collection of  $n$  arms with unknown rewards are arriving one by one in a stream, and the algorithm is only allowed to store a limited number of arms at any point. The goal is to find the arm with the largest reward while minimizing the number of arm pulls (sample complexity) and the maximum number of arms stored in the memory (space complexity). Assadi and Wang designed an algorithm for this problem that uses a memory of just one arm and still achieves the sample complexity of  $\mathcal{O}(n\Delta_{[2]}^2)$  which is  $\text{worst-case}$  optimal even for non-streaming algorithms; here  $\Delta_{[i]}$  is the gap between the rewards of the best and the  $i$ -th best arms. In this paper, we extended this line of work to stochastic MABs in the streaming model with the  $\text{instance-sensitive}$  sample complexity, i.e. the sample complexity of  $\mathcal{O}(\sum_{i=2}^n \frac{1}{\Delta_{[i]}^2} \log \log (\frac{1}{\Delta_{[i]}}))$ , similar in spirit to Karnin et.al. [ICML 2013] and Jamieson et.al. [COLT 2014] in the classical setting. We devise strong negative results under this setting: our results show that any streaming algorithm under a single pass has to use either asymptotically higher sample complexity than the instance-sensitive bound, or a memory of  $\Omega(n)$  arms, even if the parameter  $\Delta_{[2]}$  is known. In fact, the lower bound holds under much stronger assumptions, including the random order streams or a known quantity of the sample complexity. We complement our lower bounds by proposing a new algorithm that uses a memory of a single arm and achieves the instance-optimal sample complexity when all the strong assumptions hold simultaneously. Our results are developed based on a novel  $\text{arm trapping lemma}$ . This generic complexity result shows that any algorithm to  $\text{trap}$  the index of the best arm among  $o(n)$  indices (but not necessarily  $\text{find}$  it) has to use  $\Theta(n\Delta_{[2]}^2)$  sample complexity. This result is  $\text{not}$  restricted to the streaming setting, and to the best of our knowledge, this is the first result that captures the sample-space trade-off for `trapping' arms in multi-armed bandits, and it can be of independent interest.

## [Evaluation beyond Task Performance: Analyzing Concepts in AlphaZero in Hex](#)

- Charles Lovering · Jessica Forde · George Konidaris · Ellie Pavlick · Michael Littman
- abstract@[open-review](#): AlphaZero, an approach to reinforcement learning that couples neural networks and Monte Carlo tree search (MCTS), has produced state-of-the-art strategies for traditional board games like chess, Go, shogi, and Hex. While researchers and game commentators have suggested that AlphaZero uses concepts that humans consider important, it is unclear how these concepts are captured in the network. We investigate AlphaZero's internal representations in the game of Hex using two evaluation techniques from natural language processing (NLP): model probing and behavioral tests. In doing so, we introduce several new evaluation tools to the RL community, and illustrate how evaluations other than task performance can be used to provide a more complete picture of a model's strengths and weaknesses. Our analyses in the game of Hex reveal interesting patterns and generate some testable hypotheses about how such models learn in general. For example, we find that the MCTS discovers concepts before the neural network learns to encode them. We also find that concepts related to short-term end-game planning are best encoded in the final layers of the model, whereas concepts related to long-term planning are encoded in the middle layers of the model.

## [Bridging Central and Local Differential Privacy in Data Acquisition Mechanisms](#)

- Alireza Fallah · Ali Makhoul · azarakhs malekian · Asuman Ozdaglar
- abstract@[open-review](#): We study the design of optimal Bayesian data acquisition mechanisms for a platform interested in estimating the mean of a distribution by collecting data from privacy-conscious users. In our setting, users have heterogeneous sensitivities for two types of privacy losses corresponding to local and central differential privacy measures. The local privacy loss is due to the leakage of a user's information when she shares her

data with the platform, and the central privacy loss is due to the released estimate by the platform to the public. The users share their data in exchange for a payment (e.g., through monetary transfers or services) that compensates for their privacy losses. The platform does not know the privacy sensitivity of users and must design a mechanism to solicit their preferences and then deliver both local and central privacy guarantees while minimizing the estimation error plus the expected payment to users. We first establish minimax lower bounds for the estimation error, given a vector of privacy guarantees, and show that a linear estimator is (near) optimal. We then turn to our main goal: designing an optimal data acquisition mechanism. We establish that the design of such mechanisms in a Bayesian setting (where the platform knows the distribution of users' sensitivities and not their realizations) can be cast as a nonconvex optimization problem. Additionally, for the class of linear estimators, we prove that finding the optimal mechanism admits a Polynomial Time Approximation Scheme.

### Intra-agent speech permits zero-shot task acquisition

- Chen Yan Â· Federico Carnevale Â· Petko I Georgiev Â· Adam Santoro Â· Aurelia Guy Â· Alistair Muldal Â· Chia-Chun Hung Â· Joshua Abramson Â· Timothy Lillicrap Â· Gregory Wayne
- abstract@[open-review](#): Human language learners are exposed to a trickle of informative, context-sensitive language, but a flood of raw sensory data. Through both social language use and internal processes of rehearsal and practice, language learners are able to build high-level, semantic representations that explain their perceptions. Here, we take inspiration from such processes of "inner speech" in humans (Vygotsky, 1934) to better understand the role of intra-agent speech in embodied behavior. First, we formally pose intra-agent speech as a semi-supervised problem and develop two algorithms that enable visually grounded captioning with little labeled language data. We then experimentally compute scaling curves over different amounts of labeled data and compare the data efficiency against a supervised learning baseline. Finally, we incorporate intra-agent speech into an embodied, mobile manipulator agent operating in a 3D virtual world, and show that with as few as 150 additional image captions, intra-agent speech endows the agent with the ability to manipulate and answer questions about a new object without any related task-directed experience (zero-shot). Taken together, our experiments suggest that modelling intra-agent speech is effective in enabling embodied agents to learn new tasks efficiently and without direct interaction experience.

### PALMER: Perception-Action Loop with Memory Reorganization for Planning

- Onur Beker Â· Mohammad Mohammadi Â· Amir Zamir
- abstract@[open-review](#): To achieve full autonomy in a priori unknown real-world scenarios, agents should be able to operate without assuming any auxiliary instrumentation in the environment, act from high-dimensional sensory input, learn from past experience to adapt and improve, and be capable of long horizon reasoning. Classical planning algorithms are proficient at addressing the last requirement, while deep learning methods can provide the necessary flexibility to address the others. To get the best of both worlds, recent work has proposed goal-reaching methods that combine local policies and reachability estimates obtained through reinforcement learning with global graph search algorithms for long-horizon planning. However, the prevailing methods are still quite brittle, as false predictions from learning-based components can often severely debilitate downstream planning. To address this, we introduce a general-purpose planning algorithm called PALMER that creates a tight feedback loop between reinforcement learning, representation learning, sampling-based motion planning, and a non-parametric memory. Our approach augments previous work in two main ways: i) empirically, it is substantially more robust to false predictions, ii) conceptually, it defines new abstractions that can extend this line of work from goal-reaching problems to general RL problems with arbitrary reward functions, and can also offer compatibility with existing motion planning pipelines.

### Effective Dimension in Bandit Problems under Censorship

- Gauthier Guinet Â· Saurabh Amin Â· Patrick Jaillet
- abstract@[open-review](#): In this paper, we study both multi-armed and contextual bandit problems in censored environments. Our goal is to estimate the performance loss due to censorship in the context of classical algorithms designed for uncensored environments. Our main contributions include the introduction of a broad class of censorship models and their analysis in terms of the effective dimension of the problem -- a natural measure of its underlying statistical complexity and main driver of the regret bound. In particular, the effective dimension allows us to maintain the structure of the original problem at first order, while embedding it in a bigger space, and thus naturally leads to results analogous to uncensored settings. Our analysis involves a continuous generalization of the Elliptical Potential Inequality, which we believe is of independent interest. We also discover an interesting property of decision-making under censorship: a transient phase during which initial misspecification of censorship is self-corrected at an extra cost; followed by a stationary phase that reflects the inherent slowdown of learning governed by the effective dimension. Our results are useful for applications of sequential decision-making models where the feedback received depends on strategic uncertainty (e.g., agentsâ€™ willingness to follow a recommendation) and/or random uncertainty (e.g., loss or delay in arrival of information).

### LBD: Decouple Relevance and Observation for Individual-Level Unbiased Learning to Rank

- Mouxiang Chen Â· Chenghao Liu Â· Zemin Liu Â· Jianling Sun
- abstract@[open-review](#): Using Unbiased Learning to Rank (ULTR) to train the ranking model with biased click logs has attracted increased research interest. The key idea is to explicitly model the user's observation behavior when building the ranker with a large number of click logs. Considering the simplicity, recent efforts are mainly based on the position bias hypothesis, in which the observation only depends on the position. However, this hypothesis does not hold in many scenarios due to the neglect of the distinct characteristics of individuals in the same position. On the other hand, directly modeling observation bias for each individual is quite challenging, since the effects of each individual's features on relevance and observation are entangled. It is difficult to unravel this coupled effect and thus obtain a correct relevance model from click data. To address this issue, we first present the concept of coupling effect for individual-level ULTR. Then, we develop the novel Lipschitz and Bernoulli Decoupling (LBD) model to decouple the effects on relevance and observation at the individual level. We prove theoretically that our proposed method could recover the correct relevance order for the ranking objective. Empirical results on two LTR benchmark datasets show that the proposed model outperforms the state-of-the-art baselines and verify its effectiveness in debiasing data.

### Sub-exponential time Sum-of-Squares lower bounds for Principal Components Analysis

- Aaron Potechin Â· GOUTHAM RAJENDRAN
- abstract@[open-review](#): Principal Components Analysis (PCA) is a dimension-reduction technique widely used in machine learning and statistics. However, due to the dependence of the principal components on all the dimensions, the components are notoriously hard to interpret. Therefore, a variant known as sparse PCA is often preferred. Sparse PCA learns principal components of the data but enforces that such components must be sparse. This has applications in diverse fields such as computational biology and image processing. To learn sparse principal components, it's well known that standard PCA will not work, especially in high dimensions, and therefore algorithms for sparse PCA are often studied as a separate endeavor. Various algorithms have been proposed for Sparse PCA over the years, but given how fundamental it is for applications in science, the limits of efficient algorithms are only partially understood. In this work, we study the limits of the powerful Sum of Squares (SoS) family of algorithms for Sparse PCA. SoS algorithms have recently revolutionized robust statistics, leading to breakthrough algorithms for long-standing open problems in machine learning, such as optimally learning mixtures of gaussians, robust clustering, robust regression, etc. Moreover, it is believed to be the optimal robust algorithm for many statistical problems. Therefore, for sparse PCA, it's plausible that it can beat simpler algorithms such as diagonal thresholding that have been traditionally used. In this work, we show that this is not the case, by exhibiting strong tradeoffs between the number of samples required, the sparsity and the ambient dimension, for which SoS algorithms, even if allowed sub-exponential time, will fail to optimally recover the component. Our results are complemented by known algorithms in literature, thereby painting an almost complete picture of the behavior of efficient algorithms for sparse PCA. Since SoS algorithms encapsulate many algorithmic techniques such as spectral or statistical query algorithms, this solidifies the message that known algorithms are

optimal for sparse PCA. Moreover, our techniques are strong enough to obtain similar tradeoffs for Tensor PCA, another important higher order variant of PCA with applications in topic modeling, video processing, etc.

## [What Can the Neural Tangent Kernel Tell Us About Adversarial Robustness?](#)

- Nikolaos Tsilivis · Julia Kempe
- abstract@[open-review](#): Adversarial vulnerability of neural nets, and subsequent techniques to create robust models have attracted significant attention; yet we still lack a full understanding of this phenomenon. Here, we study adversarial examples of trained neural networks through analytical tools afforded by recent theory advances connecting neural networks and kernel methods, namely the Neural Tangent Kernel (NTK), following a growing body of work that keeps leveraging the NTK approximation to successfully analyze important deep learning phenomena and design algorithms for new applications. We show how NTKs allow to generate adversarial examples in a "training-free" fashion, and demonstrate that they transfer to fool their finite-width neural net counterparts in the "lazy" regime. We leverage this connection to provide an alternative view on robust and non-robust features, which have been suggested to underlie the adversarial brittleness of neural nets. Specifically, we define and study features induced by the eigendecomposition of the associated kernel to better understand the role of robust and non-robust features, the reliance on both for standard classification and the robustness-accuracy trade-off. We find that such features are surprisingly consistent across architectures, and that robust features tend to correspond to the largest eigenvalues of the model, and thus are learned early during training. Our framework allows us to identify and visualize non-robust yet useful features. Finally, we shed light on the robustness mechanism underlying adversarial training of neural nets used in practice: quantifying the evolution of the associated empirical NTK, we demonstrate that its dynamics falls much earlier into the "lazy" kernel regime and manifests a much stronger form of the well known bias to prioritize learning features within the top eigenspaces of the kernel, compared to standard training.

## [Formulating Robustness Against Unforeseen Attacks](#)

- Sihui Dai · Saeed Mahloujifar · Prateek Mittal
- abstract@[open-review](#): Existing defenses against adversarial examples such as adversarial training typically assume that the adversary will conform to a specific or known threat model, such as  $\ell_p$  perturbations within a fixed budget. In this paper, we focus on the scenario where there is a mismatch in the threat model assumed by the defense during training, and the actual capabilities of the adversary at test time. We ask the question: if the learner trains against a specific "source" threat model, when can we expect robustness to generalize to a stronger "target" threat model during test-time? Our key contribution is to formally define the problem of learning and generalization with an unforeseen adversary, which helps us reason about the increase in adversarial risk from the conventional perspective of a known adversary. Applying our framework, we derive a generalization bound which relates the generalization gap between source and target threat models to variation of the feature extractor, which measures the expected maximum difference between extracted features across a given threat model. Based on our generalization bound, we propose variation regularization (VR) which reduces variation of the feature extractor across the source threat model during training. We empirically demonstrate that using VR can lead to improved generalization to unforeseen attacks during test-time, and combining VR with perceptual adversarial training (Laidlaw et al., 2021) achieves state-of-the-art robustness on unforeseen attacks. Our code is publicly available at <https://github.com/inspire-group/variation-regularization>.

## [GRASP: Navigating Retrosynthetic Planning with Goal-driven Policy](#)

- Yemin Yu · Ying Wei · Kun Kuang · Zhengxing Huang · Huaxiu Yao · Fei Wu
- abstract@[open-review](#): Retrosynthetic planning occupies a crucial position in synthetic chemistry and, accordingly, drug discovery, which aims to find synthetic pathways of a target molecule through a sequential decision-making process on a set of feasible reactions. While the majority of recent works focus on the prediction of feasible reactions at each step, there have been limited attempts toward improving the sequential decision-making policy. Existing strategies rely on either the expensive and high-variance value estimation by online rollout, or a settled value estimation neural network pre-trained with simulated pathways of limited diversity and no negative feedback. Besides, how to return multiple candidate pathways that are not only diverse but also desirable for chemists (e.g., affordable building block materials) remains an open challenge. To this end, we propose a Goal-dRiven Actor-critic retroSynthetic Planning (GRASP) framework, where we identify the policy that performs goal-driven retrosynthesis navigation toward a user-demand objective. Our experiments on the benchmark Pistachio dataset and a chemists-designed dataset demonstrate that the framework outperforms state-of-the-art approaches by up to 32.2% on search efficiency and 5.6% on quality. Remarkably, our user studies show that GRASP successfully plans pathways that accomplish the goal prescribed with a designated goal (building block materials).

## [Retrospective Adversarial Replay for Continual Learning](#)

- Lilly Kumari · Shengjie Wang · Tianyi Zhou · Jeff A Bilmes
- abstract@[open-review](#): Continual learning is an emerging research challenge in machine learning that addresses the problem where models quickly fit the most recently trained-on data and are prone to catastrophic forgetting due to distribution shifts --- it does this by maintaining a small historical replay buffer. To avoid these problems, this paper proposes a method, ``Retrospective Adversarial Replay (RAR)'', that synthesizes adversarial samples near the forgetting boundary. RAR perturbs a buffered sample towards its nearest neighbor drawn from the current task in a latent representation space. By replaying such samples, we are able to refine the boundary between previous and current tasks, hence combating forgetting and reducing bias towards the current task. To mitigate the severity of a small replay buffer, we develop a novel MixUp-based strategy to increase replay variation by replaying mixed augmentations. Combined with RAR, this achieves a holistic framework that helps to alleviate catastrophic forgetting. We show that this excels on broadly-used benchmarks and outperforms other continual learning baselines especially when only a small buffer is used. We conduct a thorough ablation study over each key component as well as a hyperparameter sensitivity analysis to demonstrate the effectiveness and robustness of RAR.

## [Domain Adaptation under Open Set Label Shift](#)

- Saurabh Garg · Sivaraman Balakrishnan · Zachary Lipton
- abstract@[open-review](#): We introduce the problem of domain adaptation under Open Set Label Shift (OSLS), the label distribution can change arbitrarily and a new class may arrive during deployment, but the class-conditional distributions  $p(x|y)$  are domain-invariant. OSLS subsumes domain adaptation under label shift and Positive-Unlabeled (PU) learning. The learner's goals here are two-fold: (a) estimate the target label distribution, including the novel class; and (b) learn a target classifier. First, we establish necessary and sufficient conditions for identifying these quantities. Second, motivated by advances in label shift and PU learning, we propose practical methods for both tasks that leverage black-box predictors. Unlike typical Open Set Domain Adaptation (OSDA) problems, which tend to be ill-posed and amenable only to heuristics, OSLS offers a well-posed problem amenable to more principled machinery. Experiments across numerous semi-synthetic benchmarks on vision, language, and medical datasets demonstrate that our methods consistently outperform OSDA baselines, achieving  $\sim 10\%-25\%$  improvements in target domain accuracy. Finally, we analyze the proposed methods, establishing finite-sample convergence to the true label marginal and convergence to optimal classifier for linear models in a Gaussian setup. Code is available at <https://github.com/Neurips2022Anon>.

## [RKHS-SHAP: Shapley Values for Kernel Methods](#)

- Siu Lun Chau · Robert Hu · Javier González · Dino Sejdinovic
- abstract@[open-review](#): Feature attribution for kernel methods is often heuristic and not individualised for each prediction. To address this, we turn to the concept of Shapley values (SV), a coalition game theoretical framework that has previously been applied to different machine learning model interpretation tasks, such as linear models, tree ensembles and deep networks. By analysing SVs from a functional perspective, we propose RKHS-SHAP,

an attribution method for kernel machines that can efficiently compute both Interventional and Observational Shapley values using kernel mean embeddings of distributions. We show theoretically that our method is robust with respect to local perturbations - a key yet often overlooked desideratum for consistent model interpretation. Further, we propose Shapley regulariser, applicable to a general empirical risk minimisation framework, allowing learning while controlling the level of specific feature's contributions to the model. We demonstrate that the Shapley regulariser enables learning which is robust to covariate shift of a given feature and fair learning which controls the SVs of sensitive features.

## [Implicit Regularization or Implicit Conditioning? Exact Risk Trajectories of SGD in High Dimensions](#)

- Courtney Paquette · Elliot Paquette · Ben Adlam · Jeffrey Pennington
- abstract@[open-review](#): Stochastic gradient descent (SGD) is a pillar of modern machine learning, serving as the go-to optimization algorithm for a diverse array of problems. While the empirical success of SGD is often attributed to its computational efficiency and favorable generalization behavior, neither effect is well understood and disentangling them remains an open problem. Even in the simple setting of convex quadratic problems, worst-case analyses give an asymptotic convergence rate for SGD that is no better than full-batch gradient descent (GD), and the purported implicit regularization effects of SGD lack a precise explanation. In this work, we study the dynamics of multi-pass SGD on high-dimensional convex quadratics and establish an asymptotic equivalence to a stochastic differential equation, which we call homogenized stochastic gradient descent (HSGD), whose solutions we characterize explicitly in terms of a Volterra integral equation. These results yield precise formulas for the learning and risk trajectories, which reveal a mechanism of implicit conditioning that explains the efficiency of SGD relative to GD. We also prove that the noise from SGD negatively impacts generalization performance, ruling out the possibility of any type of implicit regularization in this context. Finally, we show how to adapt the HSGD formalism to include streaming SGD, which allows us to produce an exact prediction for the excess risk of multi-pass SGD relative to that of streaming SGD (bootstrap risk).

## [Exploring Linear Feature Scalability of Vision Transformer for Parameter-efficient Fine-tuning](#)

- Dongze Lian · Daquan Zhou · Jiashi Feng · Xinchao Wang
- abstract@[open-review](#): Existing fine-tuning schemes for vision transformers mostly involve updating all model parameters to adapt to a new target task. Such a practice, however, suffers from two main drawbacks: i) the original information preserved within the model is inevitably eliminated, making it unavailable for other target tasks; ii) the source dataset of the pre-training network is typically significantly larger than the target dataset, hence updating all parameters seems to be unnecessary and a waste of computational resources. In this paper, we explore a simple yet effective strategy for adapting pre-trained models, termed as LInear Feature Scalability (LIFTs). Unlike prior approaches that rely on updating the pre-trained weights of the network for downstream adaptation, LIFTs preserves 100% model parameters unaltered from the pre-trained network, thereby making the model readily reusable across a wide range of tasks. Specifically, we introduce a novel task-specific linear feature scalability adaptation layer delicately designed for each block, which takes into account the linear transformation of the features. In this way, we surprisingly show that such a linear transformation is capable of handling efficient fine-tuning. Extensive experiments on downstream tasks show the effectiveness and superior performance of our proposed LIFTs. Code will be available.

## [Online Training Through Time for Spiking Neural Networks](#)

- Mingqing Xiao · Qingyan Meng · Zongpeng Zhang · Di He · Zhouchen Lin
- abstract@[open-review](#): Spiking neural networks (SNNs) are promising brain-inspired energy-efficient models. Recent progress in training methods has enabled successful deep SNNs on large-scale tasks with low latency. Particularly, backpropagation through time (BPTT) with surrogate gradients (SG) is popularly used to enable models to achieve high performance in a very small number of time steps. However, it is at the cost of large memory consumption for training, lack of theoretical clarity for optimization, and inconsistency with the online property of biological learning rules and rules on neuromorphic hardware. Other works connect the spike representations of SNNs with equivalent artificial neural network formulation and train SNNs by gradients from equivalent mappings to ensure descent directions. But they fail to achieve low latency and are also not online. In this work, we propose online training through time (OTTT) for SNNs, which is derived from BPTT to enable forward-in-time learning by tracking presynaptic activities and leveraging instantaneous loss and gradients. Meanwhile, we theoretically analyze and prove that the gradients of OTTT can provide a similar descent direction for optimization as gradients from equivalent mapping between spike representations under both feedforward and recurrent conditions. OTTT only requires constant training memory costs agnostic to time steps, avoiding the significant memory costs of BPTT for GPU training. Furthermore, the update rule of OTTT is in the form of three-factor Hebbian learning, which could pave a path for online on-chip learning. With OTTT, it is the first time that the two mainstream supervised SNN training methods, BPTT with SG and spike representation-based training, are connected, and meanwhile it is in a biologically plausible form. Experiments on CIFAR-10, CIFAR-100, ImageNet, and CIFAR10-DVS demonstrate the superior performance of our method on large-scale static and neuromorphic datasets in a small number of time steps.

## [AutoLink: Self-supervised Learning of Human Skeletons and Object Outlines by Linking Keypoints](#)

- Xingzhe He · Bastian Wandt · Helge Rhodin
- abstract@[open-review](#): Structured representations such as keypoints are widely used in pose transfer, conditional image generation, animation, and 3D reconstruction. However, their supervised learning requires expensive annotation for each target domain. We propose a self-supervised method that learns to disentangle object structure from the appearance with a graph of 2D keypoints linked by straight edges. Both the keypoint location and their pairwise edge weights are learned, given only a collection of images depicting the same object class. The resulting graph is interpretable, for example, AutoLink recovers the human skeleton topology when applied to images showing people. Our key ingredients are i) an encoder that predicts keypoint locations in an input image, ii) a shared graph as a latent variable that links the same pairs of keypoints in every image, iii) an intermediate edge map that combines the latent graph edge weights and keypoint locations in a soft, differentiable manner, and iv) an inpainting objective on randomly masked images. Although simpler, AutoLink outperforms existing self-supervised methods on the established keypoint and pose estimation benchmarks and paves the way for structure-conditioned generative models on more diverse datasets. Project website: <https://xingzhehe.github.io/autolink/>.

## [Neural Correspondence Prior for Effective Unsupervised Shape Matching](#)

- Souhaib Attaiki · Maks Ovsjanikov
- abstract@[open-review](#): We present Neural Correspondence Prior (NCP), a new paradigm for computing correspondences between 3D shapes. Our approach is fully unsupervised and can lead to high quality correspondences even in challenging cases such as sparse point clouds or non-isometric meshes, where current methods fail. Our first key observation is that, in line with neural priors observed in other domains, recent network architectures on 3D data, even without training, tend to produce pointwise features that induce plausible maps between rigid or non-rigid shapes. Secondly, we show that given a noisy map as input, training a feature extraction network with the input map as supervision, tends to remove artifacts from the input and can act as a powerful correspondence denoising mechanism, both between individual pairs and within a collection. With these observations in hand, we propose a two-stage unsupervised paradigm for shape matching, by (i) performing unsupervised training by adapting an existing approach to obtain an initial set of noisy matches, (ii) using these matches to train a network in a supervised manner. We demonstrate that this approach significantly improves the accuracy of the maps, especially when trained within a collection. We show that NCP is data-efficient, fast, and achieves state-of-the-art results on many tasks. Our code will be released after publication.

## [LGDN: Language-Guided Denoising Network for Video-Language Modeling](#)

- Haoyu Lu · Mingyu Ding · Nanyi Fei · Yuqi Huo · Zhiwu Lu
- abstract@[open-review](#): Video-language modeling has attracted much attention with the rapid growth of web videos. Most existing methods assume that the video frames and text description are semantically correlated, and focus on video-language modeling at video level. However, this hypothesis often fails for two reasons: (1) With the rich semantics of video contents, it is difficult to cover all frames with a single video-level description; (2) A raw video typically has noisy/meaningless information (e.g., scenery shot, transition or teaser). Although a number of recent works deploy attention mechanism to alleviate this problem, the irrelevant/noisy information still makes it very difficult to address. To overcome such challenge, we thus propose an efficient and effective model, termed Language-Guided Denoising Network (LGDN), for video-language modeling. Different from most existing methods that utilize all extracted video frames, LGDN dynamically filters out the misaligned or redundant frames under the language supervision and obtains only 2--4 salient frames per video for cross-modal token-level alignment. Extensive experiments on five public datasets show that our LGDN outperforms the state-of-the-arts by large margins. We also provide detailed ablation study to reveal the critical importance of solving the noise issue, in hope of inspiring future video-language work.