Wiring Up Vision: Minimizing Supervised Synaptic Updates Needed to Produce a Primate Ventral Stream

Franziska Geiger,Martin Schrimpf,Tiago Marques,James J. DiCarlo

After training on large datasets, certain deep neural networks are surprisingly good models of the neural mechanisms of adult primate visual object recognition. Nevertheless, these models are considered poor models of the development of the visual system because they posit millions of sequential, precisely coordinated synaptic updates, each based on a labeled image. While ongoing research is pursuing the use of unsupervised proxies for labels, we here explore a complementary strategy of reducing the required number of supervised synaptic updates to produce an adult-like ventral visual stream (as judged by the match to V1, V2, V4, IT, and behavior). Such models might require less precise machinery and energy expenditure to coordinate these updates and would thus move us closer to viable neuroscientific hypotheses about how the visual system wires itself up. Relative to standard model training on labeled images in ImageNet, we here demonstrate that the total number of supervised weight updates can be substantially reduced using three complementary strategies: First, we find that only 2% of supervised updates (epochs and images) are needed to achieve 80% of the match to adult ventral stream. Specifically, training benefits predictions of higher visual cortex the most whereas early visual cortex predictions only improve marginally over the course of training. Second, by improving the random distribution of synaptic connectivity, we find that 54% of the brain match can already be achieved “at birth" (i.e. no training at all). Third, we find that, by training only 5% of model synapses, we can still achieve nearly 80% of the match to the ventral stream. This approach further improves on ImageNet performance over previous attempts in computer vision of minimizing trained components without substantially increasing the relative number of trained parameters. These results reflect first steps in modeling not just primate adult visual processing during inference, but also how the ventral visual stream might be "wired up" by evolution (a model's "birth" state) and by developmental learning (a model's updates based on visual experience).

****

Learning to Downsample for Segmentation of Ultra-High Resolution Images

Chen Jin,Ryutaro Tanno,Thomy Mertzanidou,Eleftheria Panagiotaki,Daniel C. Alexander

Many computer vision systems require low-cost segmentation algorithms based on deep learning, either because of the enormous size of input images or limited computational budget. Common solutions uniformly downsample the input images to meet memory constraints, assuming all pixels are equally informative. In this work, we demonstrate that this assumption can harm the segmentation performance because the segmentation difficulty varies spatially (see Figure 1 “Uniform”). We combat this problem by introducing a learnable downsampling module, which can be optimised together with the given segmentation model in an end-to-end fashion. We formulate the problem of training such downsampling module as optimisation of sampling density distributions over the input images given their low-resolution views. To defend against degenerate solutions (e.g. over-sampling trivial regions like the backgrounds), we propose a regularisation term that encourages the sampling locations to concentrate around the object boundaries. We find the downsampling module learns to sample more densely at difficult locations, thereby improving the segmentation performance (see Figure 1 "Ours"). Our experiments on benchmarks of high-resolution street view, aerial and medical images demonstrate substantial improvements in terms of efficiency-and-accuracy trade-off compared to both uniform downsampling and two recent advanced downsampling techniques.

****

Variational Neural Cellular Automata

Rasmus Berg Palm,Miguel González Duque,Shyam Sudhakaran,Sebastian Risi

In nature, the process of cellular growth and differentiation has lead to an amazing diversity of organisms --- algae, starfish, giant sequoia, tardigrades, and orcas are all created by the same generative process.

Inspired by the incredible diversity of this biological generative process, we propose a generative model, the Variational Neural Cellular Automata (VNCA), which is loosely inspired by the biological processes of cellular growth and differentiation. Unlike previous related works, the VNCA is a proper probabilistic generative model, and we evaluate it according to best practices. We find that the VNCA learns to reconstruct samples well and that despite its relatively few parameters and simple local-only communication, the VNCA can learn to generate a large variety of output from information encoded in a common vector format. While there is a significant gap to the current state-of-the-art in terms of generative modeling performance, we show that the VNCA can learn a purely self-organizing generative process of data. Additionally, the self-organizing nature bestows the VNCA with some inherent robustness against perturbations in the early stages of growth.

****

Wish you were here: Hindsight Goal Selection for long-horizon dexterous manipulation

Todor Davchev,Oleg Olegovich Sushkov,Jean-Baptiste Regli,Stefan Schaal,Yusuf Aytar,Markus Wulfmeier,Jon Scholz

Complex sequential tasks in continuous-control settings often require agents to successfully traverse a set of ``narrow passages'' in their state space. Solving such tasks with a sparse reward in a sample-efficient manner poses a challenge to modern reinforcement learning (RL) due to the associated long-horizon nature of the problem and the lack of sufficient positive signal during learning.

Various tools have been applied to address this challenge. When available, large sets of demonstrations can guide agent exploration. Hindsight relabelling on the other hand does not require additional sources of information. However, existing strategies explore based on task-agnostic goal distributions, which can render the solution of long-horizon tasks impractical. In this work, we extend hindsight relabelling mechanisms to guide exploration along task-specific distributions implied by a small set of successful demonstrations. We evaluate the approach on four complex, single and dual arm, robotics manipulation tasks against strong suitable baselines. The method requires far fewer demonstrations to solve all tasks and achieves a significantly higher overall performance as task complexity increases. Finally, we investigate the robustness of the proposed solution with respect to the quality of input representations and the number of demonstrations.

****

L0-Sparse Canonical Correlation Analysis

Ofir Lindenbaum,Moshe Salhov,Amir Averbuch,Yuval Kluger

Canonical Correlation Analysis (CCA) models are powerful for studying the associations between two sets of variables. The canonically correlated representations, termed \textit{canonical variates} are widely used in unsupervised learning to analyze unlabeled multi-modal registered datasets. Despite their success, CCA models may break (or overfit) if the number of variables in either of the modalities exceeds the number of samples. Moreover, often a significant fraction of the variables measures modality-specific information, and thus removing them is beneficial for identifying the \textit{canonically correlated variates}. Here, we propose ℓ_0-CCA, a method for learning correlated representations based on sparse subsets of variables from two observed modalities.

Sparsity is obtained by multiplying the input variables by stochastic gates, whose parameters are learned together with the CCA weights via an ℓ_0-regularized correlation loss.

We further propose ℓ_0-Deep CCA for solving the problem of non-linear sparse CCA by modeling the correlated representations using deep nets. We demonstrate the efficacy of the method using several synthetic and real examples. Most notably, by gating nuisance input variables, our approach improves the extracted representations compared to other linear, non-linear and sparse CCA-based models.

****

Recycling Model Updates in Federated Learning: Are Gradient Subspaces Low-Rank?

Sheikh Shams Azam,Seyyedali Hosseinalipour,Qiang Qiu,Christopher Brinton

In this paper, we question the rationale behind propagating large numbers of parameters through a distributed system during federated learning. We start by examining the rank characteristics of the subspace spanned by gradients (i.e., the gradient-space) in centralized model training, and observe that the gradient-space often consists of a few leading principal components accounting for an overwhelming majority (95-99%) of the explained variance. Motivated by this, we propose the "Look-back Gradient Multiplier" (LBGM) algorithm, which utilizes this low-rank property of the gradient-space in federated learning. Operationally, LBGM recycles the gradients between model update rounds to significantly reduce the number of parameters to be propagated through the system. We analytically characterize the convergence behavior of LBGM, revealing the nature of the trade-off between communication savings and model performance. Our subsequent experimental results demonstrate the improvement LBGM obtains on communication overhead compared to federated learning baselines. Additionally, we show that LBGM is a general plug-and-play algorithm that can be used standalone or stacked on top of existing sparsification techniques for distributed model training.

****

Generative Modeling for Multitask Visual Learning

Zhipeng Bao,Yu-Xiong Wang,Martial Hebert

Generative modeling has recently shown great promise in computer vision, but it has mostly focused on synthesizing visually realistic images. In this paper, motivated by multi-task learning of shareable feature representations, we consider a novel problem of learning a shared generative model that is useful across various visual perception tasks. Correspondingly, we propose a general multi-task oriented generative modeling (MGM) framework, by coupling a discriminative multi-task network with a generative network. While it is challenging to synthesize both RGB images and pixel-level annotations in multi-task scenarios, our framework enables us to use synthesized images paired with only weak annotations (i.e., image-level scene labels) to facilitate multiple visual tasks. Experimental evaluation on challenging multi-task benchmarks, including NYUv2 and Taskonomy, demonstrates that our MGM framework improves the performance of all the tasks by large margins, especially in the low-data regimes, and our model consistently outperforms state-of-the-art multi-task approaches.

****

Is Homophily a Necessity for Graph Neural Networks?

Yao Ma,Xiaorui Liu,Neil Shah,Jiliang Tang

Graph neural networks (GNNs) have shown great prowess in learning representations suitable for numerous graph-based machine learning tasks. When applied to semi-supervised node classification, GNNs are widely believed to work well due to the homophily assumption (``like attracts like''), and fail to generalize to heterophilous graphs where dissimilar nodes connect. Recent works design new architectures to overcome such heterophily-related limitations, citing poor baseline performance and new architecture improvements on a few heterophilous graph benchmark datasets as evidence for this notion. In our experiments, we empirically find that standard graph convolutional networks (GCNs) can actually achieve better performance than such carefully designed methods on some commonly used heterophilous graphs. This motivates us to reconsider whether homophily is truly necessary for good GNN performance. We find that this claim is not quite true, and in fact, GCNs can achieve strong performance on heterophilous graphs under certain conditions. Our work carefully characterizes these conditions and provides supporting theoretical understanding and empirical observations. Finally, we examine existing heterophilous graphs benchmarks and reconcile how the GCN (under)performs on them based on this understanding.

****

Where is the bottleneck in long-tailed classification?

Zaid Khan,Yun Fu

A commonly held belief in deep-learning based long-tailed classification is that the representations learned from long-tailed data are ”good enough” and the performance bottleneck is the classification head atop the representation learner. We design experiments to investigate this folk wisdom, and find that representation s learned from long-tailed data distributions substantially differ from the representations learned from ”normal” data distributions. We show that the long-tailed representations are volatile and brittle with respect to the true data distribution. Compared to the representations learned from the true, balanced distributions, long-tailed representations fail to localize tail classes and display vastly worse inter-class separation and intra-class compactness when unseen samples from the true data distribution are embedded into the feature space. We provide an explanation for why data augmentation helps long-tailed classification despite leaving the dataset imbalance unchanged — it promotes inter-class separation, intra-class compactness, and improves localization of tail classes w.r.t to the true data distribution.

****

Additive Poisson Process: Learning Intensity of Higher-Order Interaction in Poisson Processes

Simon Luo,Feng Zhou,lamiae azizi,Mahito Sugiyama

We present the Additive Poisson Process (APP), a novel framework that can model the higher-order interaction effects of the intensity functions in Poisson processes using projections into lower-dimensional space. Our model combines the techniques in information geometry to model higher-order interactions on a statistical manifold and in generalized additive models to use lower-dimensional projections to overcome the effects from the curse of dimensionality. Our approach solves a convex optimization problem by minimizing the KL divergence from a sample distribution in lower-dimensional projections to the distribution modeled by an intensity function in the Poisson process. Our empirical results show that our model is able to use samples observed in the lower dimensional space to estimate the higher-order intensity function with extremely sparse observations.

****

DEGREE: Decomposition Based Explanation for Graph Neural Networks

Qizhang Feng,Ninghao Liu,Fan Yang,Ruixiang Tang,Mengnan Du,Xia Hu

Graph Neural Networks (GNNs) are gaining extensive attention for their application in graph data. However, the black-box nature of GNNs prevents users from understanding and trusting the models, thus hampering their applicability. Whereas explaining GNNs remains a challenge, most existing methods fall into approximation based and perturbation based approaches with suffer from faithfulness problems and unnatural artifacts respectively. To tackle these problems, we propose DEGREE (Decomposition based Explanation for GRaph nEural nEtworks) to provide a faithful explanation for GNN predictions. By decomposing the information generation and aggregation mechanism of GNNs, DEGREE allows tracking the contributions of specific components of the input graph to the final prediction. Based on this, we further design a subgraph level interpretation algorithm to reveal complex interactions between graph nodes that are overlooked by previous methods. The efficiency of our algorithm can be further improved by utilizing GNN characteristics. Finally, we conduct quantitative and qualitative experiments on synthetic and real-world datasets to demonstrate the effectiveness of DEGREE on node classification and graph classification tasks.

****

Brain insights improve RNNs' accuracy and robustness for hierarchical control of continually learned autonomous motor motifs

Laureline Logiaco,G Sean Escola

We study the problem of learning dynamics that can produce hierarchically organized continuous outputs consisting of the flexible chaining of re-usable motor ‘motifs’ from which complex behavior is generated. Can a motif library be efficiently and extendably learned without interference between motifs, and can these motifs be chained in arbitrary orders without first learning the corresponding motif transitions during training? This requires (i) parameter updates while learning a new motif that do not interfere with the parameters used for the previously acquired ones; and (ii) successful motif generation when starting from the network states reached at the end of any of the other motifs, even if these states were not present during training (a case of out-of-distribution generalization). We meet the first requirement by designing recurrent neural networks (RNNs) with specific architectures that segregate motif-dependent parameters (as customary in continual learning works), and try a standard method to address the second by training with random initial states. We find that these standard RNNs are very unreliable during zero-shot transfer to motif chaining. We then use insights from the motor thalamocortical circuit, featuring a specific module that shapes motif transitions. We develop a method to constrain the RNNs to function similarly to the thalamocortical circuit during motif transitions, while preserving the large expressivity afforded by gradient-based training of non-analytically tractable RNNs. We then show that this thalamocortical inductive bias not only acts in synergy with gradient-descent RNN training to improve accuracy during in-training-distribution motif production, but also leads to zero-shot transfer to new motif chains with no performance cost. Besides proposing an efficient, robust and flexible RNN architecture, our results shed new light on the function of motor preparation in the brain.

****

VUT: Versatile UI Transformer for Multimodal Multi-Task User Interface Modeling

Yang Li,Gang Li,Xin Zhou,Mostafa Dehghani,Alexey A. Gritsenko

User interface modeling is inherently multimodal, which involves several distinct types of data: images, structures and language. The tasks are also diverse, including object detection, language generation and grounding. In this paper, we present VUT, a Versatile UI Transformer that takes multimodal input and simultaneously accomplishes 5 distinct tasks with the same model. Our model consists of a multimodal Transformer encoder that jointly encodes UI images and structures, and performs UI object detection when the UI structures are absent in the input. Our model also consists of an auto-regressive Transformer model that encodes the language input and decodes output, for both question-answering and command grounding with respect to the UI. Our experiments show that for most of the tasks, when trained jointly for multi-tasks, VUT has achieved accuracy either on par with or exceeding the accuracy when the model is trained for individual tasks separately.

****

DreamerPro: Reconstruction-Free Model-Based Reinforcement Learning with Prototypical Representations

Fei Deng,Ingook Jang,Sungjin Ahn

In model-based reinforcement learning (MBRL) such as Dreamer, the approaches based on observation reconstruction
often fail to discard task-irrelevant details, thus struggling to handle visual distractions or generalize to unseen distractions. To address this issue, previous work has proposed to contrastively learn the latent representations and its temporal dynamics, but showed inconsistent performance, often worse than Dreamer. Although, in computer vision, an alternative prototypical approach has often shown to be more accurate and robust, it is elusive how this approach can be combined best with the temporal dynamics learning in MBRL. In this work, we propose a reconstruction-free MBRL agent, called DreamerPro, to achieve this goal. Similar to SwAV, by encouraging uniform cluster assignment across the batch, we implicitly push apart the embeddings of different observations. Additionally, we let the temporal latent state to 'reconstruct' the cluster assignment of the observation, thereby relieving the world model from modeling low-level details. We evaluate our model on the standard setting of DeepMind Control Suite, and also on a natural background setting, where the background is replaced by natural videos irrelevant to the task. The results show that the proposed model is consistently better than the previous models.

****

Recognizing and overcoming the greedy nature of learning in multi-modal deep neural networks

Nan Wu,Stanislaw Kamil Jastrzebski,Kyunghyun Cho,Krzysztof J. Geras

We hypothesize that due to the greedy nature of learning in multi-modal deep neural networks (DNNs), these models tend to rely on just one modality while under-utilizing the other modalities. We observe empirically that such behavior hurts its overall generalization. We validate our hypothesis by estimating the gain on the accuracy when the model has access to an additional modality. We refer to this gain as the conditional utilization rate of the modality. In the experiments, we consistently observe an imbalance in conditional utilization rate between modalities, across multiple tasks and architectures. Since conditional utilization rate cannot be computed efficiently during training, we introduce an efficient proxy based on the pace at which a DNN learns from each modality, which we refer to as conditional learning speed. We thus propose a training algorithm, balanced multi-modal learning, and demonstrate that it indeed addresses the issue of greedy learning. The proposed algorithm is found to improve the model’s generalization on three datasets: Colored MNIST (Kim et al., 2019), Princeton ModelNet40 (Wu et al., 2015), and NVIDIA Dynamic Hand Gesture Dataset (Molchanov et al., 2016).

****

Subjective Learning for Open-Ended Data
Tianren Zhang,Yizhou Jiang,Xin Su,Shangqi Guo,Feng Chen

Conventional supervised learning typically assumes that the learning task can be solved by learning a single function since the data is sampled from a fixed distribution. However, this assumption is invalid in open-ended environments where no task-level data partitioning is available. In this paper, we present a novel supervised learning framework of learning from open-ended data, which is modeled as data implicitly sampled from multiple domains with the data in each domain obeying a domain-specific target function. Since different domains may possess distinct target functions, open-ended data inherently requires multiple functions to capture all its input-output relations, rendering training a single global model problematic. To address this issue, we devise an Open-ended Supervised Learning (OSL) framework, of which the key component is a subjective function that allocates the data among multiple candidate models to resolve the "conflict'' between the data from different domains, exhibiting a natural hierarchy. We theoretically analyze the learnability and the generalization error of OSL, and empirically validate its efficacy in both open-ended regression and classification tasks.

****

KINet: Keypoint Interaction Networks for Unsupervised Forward Modeling
Alireza Rezazadeh,Changhyun Choi

Object-centric representation is an essential abstraction for physical reasoning and forward prediction. Most existing approaches learn this representation through extensive supervision (e.g, object class and bounding box) although such ground-truth information is not readily accessible in reality. To address this, we introduce KINet (Keypoint Interaction Network)---an end-to-end unsupervised framework to reason about object interactions in complex systems based on a keypoint representation. Using visual observations, our model learns to associate objects with keypoint coordinates and discovers a graph representation of the system as a set of keypoint embeddings and their relations. It then learns an action-conditioned forward model using contrastive estimation to predict future keypoint states. By learning to perform physical reasoning in the keypoint space, our model automatically generalizes to scenarios with a different number of objects, and novel object geometries. Experiments demonstrate the effectiveness of our model to accurately perform forward prediction and learn plannable object-centric representations which can also be used in downstream model-based control tasks.

****

Improving Mutual Information Estimation with Annealed and Energy-Based Bounds
Rob Brekelmans,Sicong Huang,Marzyeh Ghassemi,Greg Ver Steeg,Roger Baker Grosse,Alireza Makhzani

Mutual information (MI) is a fundamental quantity in information theory and machine learning. However, direct estimation of MI is intractable, even if the true joint probability density for the variables of interest is known, as it involves estimating a potentially high-dimensional log partition function. In this work, we present a unifying view of existing MI bounds from the perspective of importance sampling, and propose three novel bounds based on this approach. Since a tight MI bound without density information requires a sample size exponential in t he true MI, we assume either a single marginal or the full joint density information is known. In settings where the full joint density is available, we propose Multi-Sample Annealed Importance Sampling (AIS) bounds on MI, which we demonstrate can tightly estimate large values of MI in our experiments. In settings where only a single marginal distribution is known, we propose Generalized IWAE (GIWAE) and MINE-AIS bounds. Our GIWAE bound unifies variational and contrastive bounds in a single framework that generalizes InfoNCE, IWAE, and Barber-Agakov bounds. Our MINE-AIS method improves upon existing energy-based methods such as MINE-DV and MINE-F by directly optimizing a tighter lower bound on MI. MINE-AIS uses MCMC sampling to estimate gradients for training and Multi-Sample AIS for evaluating the bound. Our methods are particularly suitable for evaluating MI in deep generative models, since explicit forms of the marginal or joint densities are often available. We evaluate our bounds on estimating the MI of VAEs and GANs trained on the MNIST and CIFAR datasets, and showcase significant gains over existing bounds in these challenging settings with high ground truth MI.

****

Deep banach space kernels

Mrityunjay Bhardwaj

The recent success of deep learning has encouraged many researchers to explore the deep/concatenated variants of classical kernel methods. Some of which includes MLMKL, DGP and DKL. Although, These methods have proven to be quite useful in various real-world settings. They still suffer from the limitations of only utilizing kernels from Hilbert spaces. In this paper, we address these shortcomings by introducing a new class of concatenated kernel learning methods that use the kernels from the reproducing kernel Banach spaces(RKBSs). These spaces turned out to be one of the most general spaces where a reproducing Kernel exists. We propose a framework of construction for these Deep RKBS models and then provide a representer theorem for regularized learning problems. We also describe the relationship with its deep RKHS variant as well as standard Deep Gaussian Processes. In the end, we construct and implement a two-layer deep RKBS model and demonstrate it on a range of machine learning tasks.

****

Sequence Approximation using Feedforward Spiking Neural Network for Spatiotemporal Learning: Theory and Optimization Methods

Xueyuan She,Saurabh Dash,Saibal Mukhopadhyay

A dynamical system of spiking neurons with only feedforward connections can classify spatiotemporal patterns without recurrent connections. However, the theoretical construct of a feedforward spiking neural network (SNN) for approximating a temporal sequence remains unclear, making it challenging to optimize SNN architectures for learning complex spatiotemporal patterns. In this work, we establish a theoretical framework to understand and improve sequence approximation using a feedforward SNN. Our framework shows that a feedforward SNN with one neuron per layer and skip-layer connections can approximate the mapping function between any arbitrary pairs of input and output spike train on a compact domain. Moreover, we prove that heterogeneous neurons with varying dynamics and skip-layer connections improve sequence approximation using feedforward SNN. Consequently, we propose SNN architectures incorporating the preceding constructs that are trained using supervised backpropagation-through-time (BPTT) and unsupervised spiking-timing-dependent plasticity (STDP) algorithms for classification of spatiotemporal data. A dual-search-space Bayesian optimization method is developed to optimize architecture and parameters of the proposed SNN with heterogeneous neuron dynamics and skip-layer connections.

****

Diverse Client Selection for Federated Learning via Submodular Maximization

Ravikumar Balakrishnan,Tian Li,Tianyi Zhou,Nageen Himayat,Virginia Smith,Jeff Bilmes

In every communication round of federated learning, a random subset of clients communicate their model updates back to the server which then aggregates them all. The optimal size of this subset is not known and several studies have shown that typically random selection does not perform very well in terms of convergence, learning efficiency and fairness. We, in this paper, propose to select a small diverse subset of clients, namely those carrying representative gradient information, and we transmit only these updates to the server. Our aim is for updating via only a subset to approximate updating via aggregating all client information. We achieve this by choosing a subset that maximizes a submodular facility location function defined over gradient space. We introduce “federated averaging with diverse client selection (DivFL)”. We provide a thorough analysis of its convergence in the heterogeneous setting and apply it both to synthetic and to real datasets. Empirical results show several benefits to our approach including improved learning efficiency, faster convergence and also more uniform (i.e., fair) performance across clients. We further show a communication-efficient version of DivFL that can still outperform baselines on the above metrics.

****

From Intervention to Domain Transportation: A Novel Perspective to Optimize Recommendation

Da Xu,Yuting Ye,Chuanwei Ruan,Evren Korpeoglu,Sushant Kumar,Kannan Achan

The interventional nature of recommendation has attracted increasing attention in recent years. It particularly motivates researchers to formulate learning and evaluating recommendation as causal inference and data missing-not-at-random problems. However, few take seriously the consequence of violating the critical assumption of overlapping, which we prove can significantly threaten the validity and interpretation of the outcome. We find a critical piece missing in the current understanding of information retrieval (IR) systems: as interventions, recommendation not only affects the already observed data, but it also interferes with the target domain (distribution) of interest. We then rephrase optimizing recommendation as finding an intervention that best transports the patterns it learns from the observed domain to its intervention domain. Towards this end, we use domain transportation to characterize the learning-intervention mechanism of recommendation. We design a principled transportation-constraint risk minimization objective and convert it to a two-player minimax game.

We prove the consistency, generalization, and excessive risk bounds for the proposed objective, and elaborate how they compare to the current results. Finally, we carry out extensive real-data and semi-synthetic experiments to demonstrate the advantage of our approach, and launch online testing with a real-world IR system.

****

Gesture2Vec: Clustering Gestures using Representation Learning Methods for Co-speech Gesture Generation

Payam Jome Yazdian,Mo Chen,Angelica Lim

Co-speech gestures are a principal component in conveying messages and enhancing interaction experiences between humans. Similarly, the co-speech gesture is a key ingredient in human-agent interaction including both virtual agents and robots. Existing machine learning approaches have yielded only marginal success in learning speech-to-motion at the frame level. Current methods generate repetitive gesture sequences that lack appropriateness with respect to the speech context. In this paper, we propose a Gesture2Vec model using representation learning methods to learn the relationship between semantic features and corresponding gestures. We propose a vector-quantized variational autoencoder structure as well as training techniques to learn a rigorous representation of gesture sequences. Furthermore, we use a machine translation model that takes input text and translates it into a discrete sequence of associated gesture chunks in the learned gesture space. Ultimately, we use translated quantized gestures from the input text as an input to the autoencoder’s decoder to produce gesture sequences. The resulting gestures can be applied to both virtual agents and humanoid robots. Subjective and objective evaluations confirm the success of our approach in terms of appropriateness, human-likeness, and diversity.

****

Variational Predictive Routing with Nested Subjective Timescales

Alexey Zakharov,Qinghai Guo,Zafeirios Fountas

Discovery and learning of an underlying spatiotemporal hierarchy in sequential d ata is an important topic for machine learning. Despite this, little work has been done to explore hierarchical generative models that can flexibly adapt their layerwise representations in response to datasets with different temporal dynamics. Here, we present Variational Predictive Routing (VPR) – a neural probabilistic inference system that organizes latent representations of video features in a temporal hierarchy, based on their rates of change, thus modeling continuous data as a hierarchical renewal process. By employing an event detection mechanism that relies solely on the system’s latent representations (without the need of a separate model), VPR is able to dynamically adjust its internal state following changes in the observed features, promoting an optimal organisation of representations across the levels of the model’s latent hierarchy. Using several video datasets, we show that VPR is able to detect event boundaries, disentangle spatiotemporal features across its hierarchy, adapt to the dynamics of the data, and produce accurate time-agnostic rollouts of the future. Our approach integrates insights from neuroscience and introduces a framework with high potential for applications in model-based reinforcement learning, where flexible and informative state-space rollouts are of particular interest.

****

Logit Attenuating Weight Normalization

Aman Gupta,Rohan Ramanath,Jun Shi,Anika Ramachandran,SIROU ZHU,Mingzhou Zhou,Sathiya Keerthi

Over-parameterized deep networks trained using gradient-based optimizers is a popular way of solving classification and ranking problems. Without appropriately tuned regularization, such networks have the tendency to make output scores (logits) and network weights large, causing training loss to become too small and the network to lose its adaptivity (ability to move around and escape regions of poor generalization) in the weight space. Adaptive optimizers like Adam, being aggressive at optimizing the train loss, are particularly affected by this. It is well known that, even with weight decay (WD) and normal hyper-parameter tuning, adaptive optimizers lag behind SGD a lot in terms of generalization performance, mainly in the image classification domain.

An alternative to WD for improving a network's adaptivity is to directly control the magnitude of the weights and hence the logits. We propose a method called Logit Attenuating Weight Normalization (LAWN), that can be stacked onto any gradient-based optimizer. LAWN initially starts off training in a free (unregularized) mode and, after some initial epochs, it constrains the weight norms of layers, thereby controlling the logits and improving adaptivity. This is a new regularization approach that does not use WD anywhere; instead, the number of initial free epochs becomes the new hyper-parameter. The resulting LAWN variant of adaptive optimizers gives a solid lift to generalization performance, making their performance equal or even exceed SGD's performance on benchmark image classification and recommender datasets. Another important feature is that LAWN also greatly improves the adaptive optimizers when used with large batch sizes.

****

Revisiting the Lottery Ticket Hypothesis: A Ramanujan Graph Perspective

BITHIKA PAL,Arindam Biswas,Pabitra Mitra,BISWAJIT BASU

Neural networks often yield to weight pruning resulting in a sparse subnetwork that is adequate for a given task. Retraining these `lottery ticket' subnetworks from their initialization minimizes the computational burden while preserving the test set accuracy of the original network. Based on our knowledge, the existing literature only confirms that pruning is needed and it can be achieved up to certain sparsity. We analyze the pruned network in the context of the properties of Ramanujan expander graphs. We consider the feed-forward network (both multi-layer perceptron and convolutional network) as a series of bipartite graphs which establish the connection from input to output. Now, as the fraction of remaining weights reduce with increasingly aggressive pruning two distinct regimes are observed: initially, no significant decrease in accuracy is demonstrated, and then the accuracy starts dropping rapidly. We empirically show that in the first re gime the pruned lottery ticket sub-network remains a Ramanujan graph. Subsequently, with the loss of Ramanujan graph property, accuracy begins to reduce sharply. This characterizes an absence of resilient connectivity in the pruned sub-network. We also propose a new magnitude-based pruning algorithm to preserve the above property. We perform experiments on MNIST and CIFAR10 datasets using different established feed-forward architectures and show that the winning ticket obtained from the proposed algorithm is much more robust.

****

Stochastic Projective Splitting: Solving Saddle-Point Problems with Multiple Regularizers

Patrick R. Johnstone,Jonathan Eckstein,Thomas Flynn,Shinjae Yoo

We present a new, stochastic variant of the projective splitting (PS) family of algorithms for monotone inclusion problems. It can solve min-max and noncooperative game formulations arising in applications such as robust ML without the convergence issues associated with gradient descent-ascent, the current de facto standard approach in ML applications. Our proposal is the first version of PS able to use stochastic gradient oracles. It can solve min-max games while handling multiple constraints and nonsmooth regularizers via projection and proximal operators. Unlike other stochastic splitting methods that can solve such problems, our method does not rely on a product-space reformulation of the original problem. We prove almost-sure convergence of the iterates to the solution and a convergence rate for the expected residual. By working with monotone inclusions rather than variational inequalities, our analysis avoids the drawbacks of measuring convergence through the restricted gap function. We close with numerical experiments on a distributionally robust sparse logistic regression problem.

****

RieszNet and ForestRiesz: Automatic Debiased Machine Learning with Neural Nets and Random Forests

Victor Quintas-Martinez,Victor Chernozhukov,Vasilis Syrgkanis,Whitney Newey

Many causal and policy effects of interest are defined by linear functionals of high-dimensional or non-parametric regression functions. $\sqrt{n}$-consistent and asymptotically normal estimation of the object of interest requires debiasing to reduce the effects of regularization and/or model selection on the object of interest. Debiasing is typically achieved by adding a correction term to the plug-in estimator of the functional, that is derived based on a functional-specific theoretical derivation of what is known as the influence function and which leads to properties such as double robustness and Neyman orthogonality. We instead implement an automatic debiasing procedure based on automatically learning the Riesz representation of the linear functional using Neural Nets and Random Forests. Our method solely requires value query oracle access to the linear functional. We propose a multi-tasking Neural Net debiasing method with stochastic gradient descent minimization of a combined Reisz representer and regression loss, while sharing representation layers for the two functions. We also propose a random forest method which learns a locally linear representation of the Reisz function. Even though our methodology applies to arbitrary functionals, we experimentally find that it beats state of the art performance of the prior neural net based estimator of Shi et al. (2019) for the case of the average treatment effect functional. We also evaluate our method on the more challenging problem of estimating average marginal effects with continuous treatments, using semi-synthetic data of gasoline price changes on gasoline demand.

****

New Definitions and Evaluations for Saliency Methods: Staying Intrinsic and Sound

Arushi Gupta,Nikunj Saunshi,Dingli Yu,Kaifeng Lyu,Sanjeev Arora

Saliency methods seek to provide human-interpretable explanations for the output of machine learning model on a given input. A plethora of saliency methods exist, as well as an extensive literature on their justifications/criticisms/evaluations. This paper focuses on heat maps based saliency methods that often provide explanations that look best to humans. It tries to introduce methods and evaluations for masked-based saliency methods that are {\em intrinsic} --- use just t

he training dataset and the trained net, and do not use separately trained nets, distractor distributions, human evaluations or annotations. Since a mask can be seen as a "certificate" justifying the net's answer, we introduce notions of {\em completeness} and {\em soundness} (the latter being the new contribution) motivated by logical proof systems. These notions allow a new evaluation of saliency methods, that experimentally provides a novel and stronger justification for several heuristic tricks in the field (T.V. regularization, upscaling).
****

Sample and Computation Redistribution for Efficient Face Detection
Jia Guo,Jiankang Deng,Alexandros Lattas,Stefanos Zafeiriou
Although tremendous strides have been made in uncontrolled face detection, accurate face detection with a low computation cost remains an open challenge. In this paper, we point out that computation distribution and scale augmentation are the keys to detecting small faces from low-resolution images. Motivated by these observations, we introduce two simple but effective methods: (1) Computation Redistribution (CR), which reallocates the computation between the backbone, neck and head of the model; and (2) Sample Redistribution (SR), which augments training samples for the most needed stages. The proposed Sample and Computation Redistribution for Face Detection (SCRFD) is implemented by a random search in a meticulously designed search space. Extensive experiments conducted on WIDER FACE demonstrate the state-of-the-art accuracy-efficiency trade-off for the proposed SCRFD family across a wide range of compute regimes. In particular, SCRFD-34GF outperforms the best competitor, TinaFace, by 4.78% (AP at hard set) while being more than 3$\times$ faster on GPUs with VGA-resolution images. Code is available at: https://github.com/deepinsight/insightface/tree/master/detection/scrfd.
****

Sound Adversarial Audio-Visual Navigation
Yinfeng Yu,Wenbing Huang,Fuchun Sun,Changan Chen,Yikai Wang,Xiaohong Liu
Audio-visual navigation task requires an agent to find a sound source in a realistic, unmapped 3D environment by utilizing egocentric audio-visual observations. Existing audio-visual navigation works assume a clean environment that solely contains the target sound, which, however, would not be suitable in most real-world applications due to the unexpected sound noise or intentional interference. In this work, we design an acoustically complex environment in which, besides the target sound, there exists a sound attacker playing a zero-sum game with the agent. More specifically, the attacker can move and change the volume and category of the sound to make the agent suffer from finding the sounding object while the agent tries to dodge the attack and navigate to the goal under the intervention. Under certain constraints to the attacker, we can improve the robustness of the agent towards unexpected sound attacks in audio-visual navigation. For better convergence, we develop a joint training mechanism by employing the property of a centralized critic with decentralized actors. Experiments on two real-world 3D scan datasets, Replica, and Matterport3D, verify the effectiveness and the robustness of the agent trained under our designed environment when transferred to the clean environment or the one containing sound attackers with random policy. Project: https://yyf17.github.io/SAAVN .
****

Selective Token Generation for Few-shot Language Modeling
Daejin Jo,Taehwan Kwon,Sungwoong Kim,Eun-Sol Kim
Natural language modeling with limited training data is challenging problem, and many algorithms make use of large-scale pretrained language models (PLMs) for this due to its great generalization ability. Among these transfer learning algorithms from PLMs, additive learning that incorporates a task-specific adapter on top of the fixed PLM has been popularly used to alleviate the severe overfitting problem in the few-shot setting. However, this added task-specific adapter is generally trained by maximum likelihood estimation that can easily suffer from the so-called exposure bias problem, especially in sequential text generation. Therefore, in this work, we develop a novel additive learning algorithm based on reinforcement learning (RL) for few-shot natural language generation (NLG) tasks. In particular, we propose to use a selective token generation between the transf

ormer-based PLM and the task-specific adapter during both training and inference. This output token selection between the two generators allows the adapter to take into account only on the task-relevant parts in sequence generation, and therefore makes it more robust to overfitting as well as more stable in RL training. In addition, in order to obtain the complementary adapter from the PLM for each few-shot task, we exploit a separate selecting module that is also simultaneously trained using RL. Experimental results on various few-shot NLG tasks including data-to-text generation and text summarization demonstrate that the proposed selective token generation significantly outperforms the previous additive learning algorithms based on the PLMs.

****

Out-of-distribution Generalization in the Presence of Nuisance-Induced Spurious Correlations

Aahlad Manas Puli,Lily H Zhang,Eric Karl Oermann,Rajesh Ranganath

In many prediction problems, spurious correlations are induced by a changing relationship between the label and a nuisance variable that is also correlated with the covariates. For example, in classifying animals in natural images, the background, which is a nuisance, can predict the type of animal. This nuisance-label relationship does not always hold, and the performance of a model trained under one such relationship may be poor on data with a different nuisance-label relationship. To build predictive models that perform well regardless of the nuisance-label relationship, we develop Nuisance-Randomized Distillation (NURD). We introduce the nuisance-randomized distribution, a distribution where the nuisance and the label are independent. Under this distribution, we define the set of representations such that conditioning on any member, the nuisance and the label remain independent. We prove that the representations in this set always perform better than chance, while representations outside of this set may not. NURD finds a representation from this set that is most informative of the label under the nuisance-randomized distribution, and we prove that this representation achieves the highest performance regardless of the nuisance-label relationship. We evaluate NURD on several tasks including chest X-ray classification where, using non-lung patches as the nuisance, NURD produces models that predict pneumonia under strong spurious correlations.

****

Dynamics-Aware Comparison of Learned Reward Functions

Blake Wulfe,Logan Michael Ellis,Jean Mercat,Rowan Thomas McAllister,Adrien Gaidon

The ability to learn reward functions plays an important role in enabling the deployment of intelligent agents in the real world. However, $\textit{comparing}$ reward functions, for example as a means of evaluating reward learning methods, presents a challenge. Reward functions are typically compared by considering the behavior of optimized policies, but this approach conflates deficiencies in the reward function with those of the policy search algorithm used to optimize it. To address this challenge, Gleave et al. (2020) propose the Equivalent-Policy Invariant Comparison (EPIC) distance. EPIC avoids policy optimization, but in doing so requires computing reward values at transitions that may be impossible under the system dynamics. This is problematic for learned reward functions because it entails evaluating them outside of their training distribution, resulting in inaccurate reward values that we show can render EPIC ineffective at comparing rewards. To address this problem, we propose the Dynamics-Aware Reward Distance (DARD), a new reward pseudometric. DARD uses an approximate transition model of the environment to transform reward functions into a form that allows for comparisons that are invariant to reward shaping while only evaluating reward functions on transitions close to their training distribution. Experiments in simulated physical domains demonstrate that DARD enables reliable reward comparisons without policy optimization and is significantly more predictive than baseline methods of downstream policy performance when dealing with learned reward functions.

****

AEVA: Black-box Backdoor Detection Using Adversarial Extreme Value Analysis

Junfeng Guo,Ang Li,Cong Liu

Deep neural networks (DNNs) are proved to be vulnerable against backdoor attacks. A backdoor could be embedded in the target DNNs through injecting a backdoor trigger into the training examples, which can cause the target DNNs misclassify an input attached with the backdoor trigger. Recent backdoor detection methods often require the access to the original poisoned training data, the parameters of the target DNNs, or the predictive confidence for each given input, which are impractical in many real-world applications, e.g., on-device de-ployed DNNs. We address the black-box hard-label backdoor detection problem where the DNN is a fully black-box and only its final output label is accessible. We approach this problem from the optimization perspective and show that the objective of backdoor detection is bounded by an adversarial objective. Further theoretical and empirical studies reveal that this adversarial objective leads to a solution with highly skewed distribution; a singularity is often observed in the adversarial map of a backdoor-infected example, which we call the adversarial singularity phenomenon. Based on this observation, we propose the adversarial extreme value analysis(AEVA) algorithm to detect backdoors in black-box neural networks. The AEVA algorithm is based on an extreme value analysis on the adversarial map, computed from the monte-carlo gradient estimation due to the black-box hard-label constraint. Evidenced by extensive experiments across three popular tasks and backdoor attacks, our approach is shown effective in detecting backdoor attacks under the black-box hard-label scenarios

****

Adversarial Distributions Against Out-of-Distribution Detectors

Sangwoong Yoon,Jinwon Choi,Yonghyeon LEE,Yung-Kyun Noh,Frank C. Park

Out-of-distribution (OOD) detection is the task of determining whether an input lies outside the training data distribution. As an outlier may deviate from the training distribution in unexpected ways, an ideal OOD detector should be able to detect all types of outliers. However, current evaluation protocols test a detector over OOD datasets that cover only a small fraction of all possible outliers, leading to overly optimistic views of OOD detector performance. In this paper, we propose a novel evaluation framework for OOD detection that tests a detector over a larger, unexplored space of outliers. In our framework, a detector is evaluated with samples from its adversarial distribution, which generates diverse outlier samples that are likely to be misclassified as in-distribution by the detector. Using adversarial distributions, we investigate OOD detectors with reported near-perfect performance on standard benchmarks like CIFAR-10 vs SVHN. Our methods discover a wide range of samples that are obviously outlier but recognized as in-distribution by the detectors, indicating that current state-of-the-art detectors are not as perfect as they seem on existing benchmarks.

****

Resonance in Weight Space: Covariate Shift Can Drive Divergence of SGD with Momentum

Kirby Banman,Garnet Liam Peet-Pare,Nidhi Hegde,Alona Fyshe,Martha White

Most convergence guarantees for stochastic gradient descent with momentum (SGDm) rely on iid sampling. Yet, SGDm is often used outside this regime, in settings with temporally correlated input samples such as continual learning and reinforcement learning. Existing work has shown that SGDm with a decaying step-size can converge under Markovian temporal correlation. In this work, we show that SGDm under covariate shift with a fixed step-size can be unstable and diverge. In particular, we show SGDm under covariate shift is a parametric oscillator, and so can suffer from a phenomenon known as resonance. We approximate the learning system as a time varying system of ordinary differential equations, and leverage existing theory to characterize the system's divergence/convergence as resonant/nonresonant modes. The theoretical result is limited to the linear setting with periodic covariate shift, so we empirically supplement this result to show that resonance phenomena persist even under non-periodic covariate shift, nonlinear dynamics with neural networks, and optimizers other than SGDm.

****

On the Implicit Biases of Architecture & Gradient Descent

Jeremy Bernstein,Yisong Yue

Do neural networks generalise because of bias in the functions returned by gradient descent, or bias already present in the network architecture? $\textit{¿Por qué no los dos?}$ This paper finds that while typical networks that fit the training data already generalise fairly well, gradient descent can further improve generalisation by selecting networks with a large margin. This conclusion is based on a careful study of the behaviour of infinite width networks trained by Bayesian inference and finite width networks trained by gradient descent. To measure the implicit bias of architecture, new technical tools are developed to both $\textit{analytically bound}$ and $\textit{consistently estimate}$ the average test error of the neural network--Gaussian process (NNGP) posterior. This error is found to be already better than chance, corroborating the findings of Valle-Pérez et al. (2019) and underscoring the importance of architecture. Going beyond this result, this paper finds that test performance can be substantially improved by selecting a function with much larger margin than is typical under the NNGP posterior. This highlights a curious fact: $\textit{minimum a posteriori}$ functions can generalise best, and gradient descent can select for those functions. In summary, new technical tools suggest a nuanced portrait of generalisation involving both the implicit biases of architecture and gradient descent.

****

PACE: A Parallelizable Computation Encoder for Directed Acyclic Graphs

Zehao Dong,Muhan Zhang,Fuhai Li,Yixin Chen

Optimization of directed acyclic graph (DAG) structures has many applications, such as neural architecture search (NAS) and probabilistic graphical model learning. Encoding DAGs into real vectors is a dominant component in most neural-network-based DAG optimization frameworks. Currently, most popular DAG encoders use an asynchronous message passing scheme which sequentially processes nodes according to the dependency between nodes in a DAG. That is, a node must not be processed until all its predecessors are processed. As a result, they are inherently not parallelizable. In this work, we propose a Parallelizable Attention-based Computation structure Encoder (PACE) that processes nodes simultaneously and encodes DAGs in parallel. We demonstrate the superiority of PACE through encoder-dependent optimization subroutines that search the optimal DAG structure based on the learned DAG embeddings. Experiments show that PACE not only improves the effectiveness over previous sequential DAG encoders with a significantly boosted training and inference speed, but also generates smooth latent (DAG encoding) spaces that are beneficial to downstream optimization subroutines.

****

Proper Straight-Through Estimator: Breaking symmetry promotes convergence to true minimum

Shinya Gongyo,Kohta Ishikawa

In the quantized network, its gradient shows either vanishing or diverging. The network thus cannot be learned by the standard back-propagation, so that an alternative approach called Straight Through Estimator (STE), which replaces the part of the gradient with a simple differentiable function, is used. While STE is known to work well for learning the quantized network empirically, it has not been established theoretically. A recent study by Yin et. al. (2019) has provided theoretical support for STE. However, its justification is still limited to the model in the one-hidden layer network with the binary activation where Gaussian generates the input data, and the true labels are output from the teacher network with the same binary network architecture. In this paper, we discuss the effectiveness of STEs in more general situations without assuming the shape of the input distribution and the labels. By considering the scale symmetry of the network and specific properties of the STEs, we find that STE with clipped Relu is superior to STEs with identity function and vanilla Relu. The clipped Relu STE, which breaks the scale symmetry, may pick up one of the local minima degenerated in scales, while the identity STE and vanilla Relu STE, which keep the scale symmetry, may not pick it up. To confirm this observation, we further present an analysis of a simple misspecified model as an example. We find that all the stationary points are identical with the vanishing points of the cRelu STE gradient, while some of them are not identical with the vanishing points of the identity and

Relu STE.

****

Ask2Mask: Guided Data Selection for Masked Speech Modeling

Murali Karthick Baskar,Andrew Rosenberg,Bhuvana Ramabhadran,Yu Zhang,Pedro Moreno

Masked speech modeling (MSM) methods such as wav2vec2 or w2v-BERT learn representations over speech frames which are randomly masked within an utterance. While these methods improve performance of Automatic Speech Recognition (ASR) systems, they have one major limitation. They treat all unsupervised speech samples with equal weight, which hinders learning as not all samples have relevant information to learn meaningful representations. In this work, we address this limitation. We propose ask2mask (ATM), a novel approach to focus on specific samples during MSM pre-training. ATM employs an external ASR model or \textit{scorer} to weight unsupervised input samples in two different ways: 1) A fine-grained data selection is performed by masking over the highly confident input frames as chosen by the scorer. This allows the model to learn meaningful representations. 2) ATM is further extended to focus at utterance-level by weighting the final MSM loss with the utterance-level confidence score. We conduct fine-tuning experiments on two well-benchmarked corpora: LibriSpeech (matching the pre-training data) and AMI (not matching the pre-training data). The results substantiate the efficacy of ATM on significantly improving the recognition performance under mismatched conditions (up to 11.6\% relative) while still yielding modest improvements under matched conditions.

****

Boosting the Confidence of Near-Tight Generalization Bounds for Uniformly Stable Randomized Algorithms

Xiaotong Yuan,Ping Li

High probability generalization bounds of uniformly stable learning algorithms have recently been actively studied with a series of near-tight results established by~\citet{feldman2019high,bousquet2020sharper}. However, for randomized algorithms with on-average uniform stability, such as stochastic gradient descent (SGD) with time decaying learning rates, it still remains less well understood if these deviation bounds still hold with high confidence over the internal randomness of algorithm. This paper addresses this open question and makes progress towards answering it inside a classic framework of confidence-boosting. To this end, we first establish an in-expectation first moment generalization error bound for randomized learning algorithm with on-average uniform stability, based on which we then show that a properly designed subbagging process leads to near-tight high probability generalization bounds over the randomness of data and algorithm. We further substantialize these generic results to SGD to derive improved high probability generalization bounds for convex or non-convex optimization with natural time decaying learning rates, which have not been possible to prove with the existing uniform stability results. Specially for deterministic uniformly stable algorithms, our confidence-boosting results improve upon the best known generalization bounds in terms of a logarithmic factor on sample size, which moves a step forward towards resolving an open question raised by~\citet{bousquet2020sharper}.

****

Domino: Discovering Systematic Errors with Cross-Modal Embeddings

Sabri Eyuboglu,Maya Varma,Khaled Kamal Saab,Jean-Benoit Delbrouck,Christopher Lee-Messer,Jared Dunnmon,James Zou,Christopher Re

Machine learning models that achieve high overall accuracy often make systematic errors on important subsets (or slices) of data. Identifying underperforming slices is particularly challenging when working with high-dimensional inputs (e.g. images, audio), where important slices are often unlabeled. In order to address this issue, recent studies have proposed automated slice discovery methods (SDMs), which leverage learned model representations to mine input data for slices on which a model performs poorly. To be useful to a practitioner, these methods must identify slices that are both underperforming and coherent (i.e. united by a human-understandable concept). However, no quantitative evaluation framework cu rrently exists for rigorously assessing SDMs with respect to these criteria. Additionally, prior qualitative evaluations have shown that SDMs often identify slices that are incoherent. In this work, we address these challenges by first designing a principled evaluation framework that enables a quantitative comparison of SDMs across 1,235 slice discovery settings in three input domains (natural images, medical images, and time-series data).

Then, motivated by the recent development of powerful cross-modal representation learning approaches, we present Domino, an SDM that leverages cross-modal embeddings and a novel error-aware mixture model to discover and describe coherent slices. We find that Domino accurately identifies 36% of the 1,235 slices in our framework -- a 12 percentage point improvement over prior methods. Further, Domino is the first SDM that can provide natural language descriptions of identified slices, correctly generating the exact name of the slice in 35% of settings.

****

Top-label calibration and multiclass-to-binary reductions

Chirag Gupta,Aaditya Ramdas

We propose a new notion of multiclass calibration called top-label calibration. A classifier is said to be top-label calibrated if the reported probability for the predicted class label---the top-label---is calibrated, conditioned on the top-label. This conditioning is essential for practical utility of the calibration property, since the top-label is always reported and we must condition on what is reported. However, the popular notion of confidence calibration erroneously skips this conditioning. Furthermore, we outline a multiclass-to-binary (M2B) reduction framework that unifies confidence, top-label, and class-wise calibration, among others. As its name suggests, M2B works by reducing multiclass calibration to different binary calibration problems; various types of multiclass calibration can then be achieved using simple binary calibration routines. We instantiate the M2B framework with the well-studied histogram binning (HB) binary calibrator, and prove that the overall procedure is multiclass calibrated without making any assumptions on the underlying data distribution. In an empirical evaluation with four deep net architectures on CIFAR-10 and CIFAR-100, we find that the M2B + HB procedure achieves lower top-label and class-wise calibration error than other approaches such as temperature scaling. Code for this work is available at https://github.com/aigen/df-posthoc-calibration.

****

How to Adapt Your Large-Scale Vision-and-Language Model

Konwoo Kim,Michael Laskin,Igor Mordatch,Deepak Pathak

Pre-training large-scale vision and language models (e.g. CLIP) has shown promising results in representation and transfer learning. We investigate the question of how to efficiently adapt these models to downstream tasks. For image classification, linear probes have been the standard for ease of use and efficiency, while for language, other approaches like prompt tuning have emerged. We analyze several fine-tuning methods across a diverse set of image classification tasks across two spectra investigating the amount and similarity of downstream data to that of pretraining one. We find that just tuning LayerNorm parameters is a surprisingly effective baseline across the board. We further demonstrate a simple yet effective strategy that combines LayerNorm-tuning with general fine-tuning methods to improve their performance and benchmark them on few-shot adaption and distribution shift tasks. Finally, we provide an empirical analysis and recommend general recipes for efficient transfer learning of vision and language models. Website at https://sites.google.com/view/adapt-large-scale-models

****

Anisotropic Random Feature Regression in High Dimensions

Gabriel Mel,Jeffrey Pennington

In contrast to standard statistical wisdom, modern learning algorithms typically find their best performance in the overparameterized regime in which the model has many more parameters than needed to fit the training data. A growing number of recent works have shown that random feature models can offer a detailed theoretical explanation for this unexpected behavior, but typically these analyses have utilized isotropic distributional assumptions on the underlying data generati on process, thereby failing to provide a realistic characterization of real-world models that are designed to identify and harness the structure in natural data. In this work, we examine the high-dimensional asymptotics of random feature regression in the presence of structured data, allowing for arbitrary input correlations and arbitrary alignment between the data and the weights of the target function. We define a partial order on the space of weight-data alignments and prove that generalization performance improves in response to stronger alignment. We also clarify several previous observations in the literature by distinguishing the behavior of the sample-wise and parameter-wise learning curves, finding that sample-wise multiple descent can occur at scales dictated by the eigenstructure of the data covariance, but that parameter-wise multiple descent is limited to double descent, although strong anisotropy can induce additional signatures such as wide plateaus and steep cliffs. Finally, these signatures are related to phase transitions in the spectrum of the feature kernel matrix, and unlike the double descent peak, persist even under optimal regularization.

****

Back2Future: Leveraging Backfill Dynamics for Improving Real-time Predictions in Future

Harshavardhan Kamarthi,Alexander Rodríguez,B. Aditya Prakash

For real-time forecasting in domains like public health and macroeconomics, data collection is a non-trivial and demanding task. Often after being initially released, it undergoes several revisions later (maybe due to human or technical constraints) - as a result, it may take weeks until the data reaches a stable value. This so-called ‘backfill’ phenomenon and its effect on model performance have been barely addressed in the prior literature. In this paper, we introduce the multi-variate backfill problem using COVID-19 as the motivating example.

We construct a detailed dataset composed of relevant signals over the past year of the pandemic.

We then systematically characterize several patterns in backfill dynamics and leverage our observations for formulating a novel problem and neural framework, Back2Future, that aims to refines a given model's predictions in real-time. Our extensive experiments demonstrate that our method refines the performance of the diverse set of top models for COVID-19 forecasting and GDP growth forecasting. Specifically, we show that Back2Future refined top COVID-19 models by 6.65% to 11.24% and yield an 18% improvement over non-trivial baselines. In addition, we show that our model improves model evaluation too; hence policy-makers can better understand the true accuracy of forecasting models in real-time.

****

Fast Finite Width Neural Tangent Kernel

Roman Novak,Jascha Sohl-Dickstein,Samuel Stern Schoenholz

The Neural Tangent Kernel (NTK), defined as the outer product of the neural network (NN) Jacobians, $\Theta_\theta(x_1, x_2) = \left[\partial f(\theta, x_1)\big/\partial \theta\right] \left[\partial f(\theta, x_2)\big/\partial \theta\right]^T$, has emerged as a central object of study in deep learning. In the infinite width limit, the NTK can sometimes be computed analytically and is useful for understanding training and generalization of NN architectures. At finite widths, the NTK is also used to better initialize NNs, compare the conditioning across models, perform architecture search, and do meta-learning. Unfortunately, the finite-width NTK is notoriously expensive to compute, which severely limits its practical utility.

We perform the first in-depth analysis of the compute and memory requirements for NTK computation in finite width networks.

Leveraging the structure of neural networks, we further propose two novel algorithms that change the exponent of the compute and memory requirements of the finite width NTK, dramatically improving efficiency.

We open-source (https://github.com/iclr2022anon/fast_finite_width_ntk) our two algorithms as general-purpose JAX function transformations that apply to any differentiable computation (convolutions, attention, recurrence, etc.) and introduce no new hyper-parameters.

****

SpanDrop: Simple and Effective Counterfactual Learning for Long Sequences

Peng Qi,Guangtao Wang,Jing Huang

Distilling supervision signal from a long sequence to make predictions is a challenging task in machine learning, especially when not all elements in the input sequence contribute equally to the desired output. In this paper, we propose SpanDrop, a simple and effective data augmentation technique that helps models identify the true supervision signal in a long sequence with very few examples. By directly manipulating the input sequence, SpanDrop randomly ablates parts of the sequence at a time and ask the model to perform the same task to emulate counterfactual learning and achieve input attribution. Based on theoretical analysis of its properties, we also propose a variant of SpanDrop based on the beta-Bernoulli distribution, which yields diverse augmented sequences while providing a learning objective that is more consistent with the original dataset. We demonstrate the effectiveness of SpanDrop on a set of carefully designed toy tasks, as well as various natural language processing tasks that require reasoning over long sequences to arrive at the correct answer, and show that it helps models improve performance both when data is scarce and abundant.

****

Embedding models through the lens of Stable Coloring

Aditya Desai,Shashank Sonkar,Anshumali Shrivastava,Richard Baraniuk

Embedding-based approaches find the semantic meaning of tokens in structured data such as natural language, graphs, and even images. To a great degree, these approaches have developed independently in different domains. However, we find a common principle underlying these formulations, and it is rooted in solutions to the stable coloring problem in graphs (Weisfeiler-Lehman isomorphism test). For instance, we find links between stable coloring, distribution hypothesis in natural language processing, and non-local-means denoising algorithm in image signal processing. We even find that stable coloring has strong connections to a broad class of unsupervised embedding models which is surprising at first since stable coloring is generally applied for combinatorial problems. To establish this connection concretely we define a mathematical framework that defines continuous stable coloring on graphs and develops optimization problems to search for them. Grounded on this framework, we show that many algorithms ranging across different domains are, in fact, searching for continuous stable coloring solutions of an underlying graph corresponding to the domain. We show that popular and widely used embedding models such as Word2Vec, AWE, BERT, Node2Vec, and Vis-Transformer can be understood as instantiations of our general algorithm that solves the problem of continuous stable coloring. These instantiations offer useful insights into the workings of state-of-the-art models like BERT stimulating new research directions.

****

Approximation and Learning with Deep Convolutional Models: a Kernel Perspective

Alberto Bietti

The empirical success of deep convolutional networks on tasks involving high-dimensional data such as images or audio suggests that they can efficiently approximate certain functions that are well-suited for such tasks. In this paper, we study this through the lens of kernel methods, by considering simple hierarchical kernels with two or three convolution and pooling layers, inspired by convolutional kernel networks. These achieve good empirical performance on standard vision datasets, while providing a precise description of their functional space that yields new insights on their inductive bias. We show that the RKHS consists of additive models of interaction terms between patches, and that its norm encourages spatial similarities between these terms through pooling layers. We then provide generalization bounds which illustrate how pooling and patches yield improved sample complexity guarantees when the target function presents such regularities.

****

Value Function Spaces: Skill-Centric State Abstractions for Long-Horizon Reasoning
Dhruv Shah,Peng Xu,Yao Lu,Ted Xiao,Alexander T Toshev,Sergey Levine,brian ichter
Reinforcement learning can train policies that effectively perform complex tasks. However for long-horizon tasks, the performance of these methods degrades with horizon, often necessitating reasoning over and chaining lower-level skills. Hierarchical reinforcement learning aims to enable this by providing a bank of low-level skills as action abstractions. Hierarchies can further improve on this by abstracting the space states as well. We posit that a suitable state abstraction should depend on the capabilities of the available lower-level policies. We propose Value Function Spaces: a simple approach that produces such a representation by using the value functions corresponding to each lower-level skill. These value functions capture the affordances of the scene, thus forming a representation that compactly abstracts task relevant information and robustly ignores distractors. Empirical evaluations for maze-solving and robotic manipulation tasks demonstrate that our approach improves long-horizon performance and enables better zero-shot generalization than alternative model-free and model-based methods.
****
Tesseract: Gradient Flip Score to Secure Federated Learning against Model Poisoning Attacks
Atul Sharma,Wei Chen,Joshua Christian Zhao,Qiang Qiu,Somali Chaterji,Saurabh Bagchi
Federated learning—multi-party, distributed learning in a decentralized environment—is vulnerable to model poisoning attacks, even more so than centralized learning approaches. This is because malicious clients can collude and send in carefully tailored model updates to make the global model inaccurate. This motivated the development of Byzantine-resilient federated learning algorithms, such as Krum, Trimmed mean, and FoolsGold. However, a recently developed targeted model poisoning attack showed that all prior defenses can be bypassed. The attack uses the intuition that simply by changing the sign of the gradient updates that the optimizer is computing, for a set of malicious clients, a model can be pushed away from the optima to increase the test error rate. In this work, we develop tesseract—a defense against this directed deviation attack, a state-of-the-art model poisoning attack. TESSERACT is based on a simple intuition that in a federated learning setting, certain patterns of gradient flips are indicative of an attack. This intuition is remarkably stable across different learning algorithms, models, and datasets. TESSERACT assigns reputation scores to the participating clients based on their behavior during the training phase and then takes a weighted contribution of the clients. We show that TESSERACT provides robustness against even an adaptive white-box version of the attack.
****
CareGraph: A Graph-based Recommender System for Diabetes Self-Care
Sirinart Tangruamsub,Karthik Kappaganthu,John O'Donovan,Anmol Madan
In this work, we build a knowledge graph that captures key attributes of content and notifications in a digital health platform for diabetes management. We propose a Deep Neural Network-based recommender that uses the knowledge graph embeddings to recommend health nudges for maximizing engagement by combating the cold-start and sparsity problems. We use a leave-one-out approach to evaluate the model. We compare the proposed model performance with a text similarity and Deep-and-Cross Network-based approach as the baseline. The overall improvement in Click-Through-Rate prediction AUC for the Knowledge-Graph-based model was 11%. We also observe that our model improved the average AUC by 5% in cold-start situations.
****
Natural Language Descriptions of Deep Visual Features
Evan Hernandez,Sarah Schwettmann,David Bau,Teona Bagashvili,Antonio Torralba,Jacob Andreas
Some neurons in deep networks specialize in recognizing highly specific perceptual, structural, or semantic features of inputs. In computer vision, techniques exist for identifying neurons that respond to individual concept categories like

colors, textures, and object classes. But these techniques are limited in scope, labeling only a small subset of neurons and behaviors in any network. Is a richer characterization of neuron-level computation possible? We introduce a procedure (called MILAN, for mutual information-guided linguistic annotation of neurons) that automatically labels neurons with open-ended, compositional, natural language descriptions. Given a neuron, MILAN generates a description by searching for a natural language string that maximizes pointwise mutual information with the image regions in which the neuron is active. MILAN produces fine-grained descriptions that capture categorical, relational, and logical structure in learned features. These descriptions obtain high agreement with human-generated feature descriptions across a diverse set of model architectures and tasks, and can aid in understanding and controlling learned models. We highlight three applications of natural language neuron descriptions. First, we use MILAN for analysis, characterizing the distribution and importance of neurons selective for attribute, category, and relational information in vision models. Second, we use MILAN for auditing, surfacing neurons sensitive to human faces in datasets designed to obscure them. Finally, we use MILAN for editing, improving robustness in an image classifier by deleting neurons sensitive to text features spuriously correlated with class labels.

****

Self-Supervise, Refine, Repeat: Improving Unsupervised Anomaly Detection
Jinsung Yoon,Kihyuk Sohn,Chun-Liang Li,Sercan O Arik,Chen-Yu Lee,Tomas Pfister

Anomaly detection (AD) - separating anomalies from normal data - has many applications across domains, from manufacturing to healthcare. While most previous works have been shown to be effective for cases with fully or partially labeled data, that setting is in practice less common due to labeling being particularly tedious for this task. In this paper, we focus on fully unsupervised AD, in which the entire training dataset, containing both normal and anomalous samples, is unlabeled. To tackle this problem effectively, we propose to improve the robustness of one-class classification trained on self-supervised representations using a data refinement process. Our proposed data refinement approach is based on an ensemble of one-class classifiers (OCCs), each of which is trained on a disjoint subset of training data. Representations learned by self-supervised learning on the refined data are iteratively updated as the refinement improves. We demonstrate our method on various unsupervised AD tasks with image and tabular data. With a 10% anomaly ratio on CIFAR-10 image data / 2.5% anomaly ratio on Thyroid tabular data, the proposed method outperforms the state-of-the-art one-class classification method by 6.3 AUC and 12.5 average precision / 22.9 F1-score.

****

DNBP: Differentiable Nonparametric Belief Propagation
Anthony Opipari,Jana Pavlasek,Chao Chen,Shoutian Wang,Karthik Desingh,Odest Jenkins

We present a differentiable approach to learn the probabilistic factors used for inference by a nonparametric belief propagation algorithm. Existing nonparametric belief propagation methods rely on domain-specific features encoded in the probabilistic factors of a graphical model. In this work, we replace each crafted factor with a differentiable neural network enabling the factors to be learned using an efficient optimization routine from labeled data. By combining differentiable neural networks with an efficient belief propagation algorithm, our method learns to maintain a set of marginal posterior samples using end-to-end training. We evaluate our differentiable nonparametric belief propagation (DNBP) method on a set of articulated pose tracking tasks and compare performance with learned baselines. Results from these experiments demonstrate the effectiveness of using learned factors for tracking and suggest the practical advantage over hand-crafted approaches. The project webpage is available at: https://sites.google.com/view/diff-nbp

****

Learning Hierarchical Structures with Differentiable Nondeterministic Stacks
Brian DuSell,David Chiang

Learning hierarchical structures in sequential data -- from simple algorithmic p atterns to natural language -- in a reliable, generalizable way remains a challenging problem for neural language models. Past work has shown that recurrent neural networks (RNNs) struggle to generalize on held-out algorithmic or syntactic patterns without supervision or some inductive bias. To remedy this, many papers have explored augmenting RNNs with various differentiable stacks, by analogy with finite automata and pushdown automata (PDAs). In this paper, we improve the performance of our recently proposed Nondeterministic Stack RNN (NS-RNN), which uses a differentiable data structure that simulates a nondeterministic PDA, with two important changes. First, the model now assigns unnormalized positive weights instead of probabilities to stack actions, and we provide an analysis of why this improves training. Second, the model can directly observe the state of the underlying PDA. Our model achieves lower cross-entropy than all previous stack RNNs on five context-free language modeling tasks (within 0.05 nats of the information-theoretic lower bound), including a task on which the NS-RNN previously failed to outperform a deterministic stack RNN baseline. Finally, we propose a restricted version of the NS-RNN that incrementally processes infinitely long sequences, and we present language modeling results on the Penn Treebank.

****

Fast Regression for Structured Inputs

Raphael A Meyer,Cameron N Musco,Christopher P Musco,David Woodruff,Samson Zhou

We study the ℓ_p regression problem, which requires finding $\mathbf{x}\in\mathbb R^{d}$ that minimizes $\|\mathbf{A}\mathbf{x}-\mathbf{b}\|_p$ for a matrix $\mathbf{A}\in\mathbb R^{n \times d}$ and response vector $\mathbf{b}\in\mathbb R^{n}$. There has been recent interest in developing subsampling methods for this problem that can outperform standard techniques when n is very large. However, all known subsampling approaches have run time that depends exponentially on p, typically, $d^{\mathcal{O}(p)}$, which can be prohibitively expensive.

We improve on this work by showing that for a large class of common \emph{structured matrices}, such as combinations of low-rank matrices, sparse matrices, and Vandermonde matrices, there are subsampling based methods for ℓ_p regression that depend polynomially on p. For example, we give an algorithm for ℓ_p regression on Vandermonde matrices that runs in time $\mathcal{O}(n\log^3 n+(dp^2)^{0.5+\omega}\cdot\text{polylog}\,n)$, where ω is the exponent of matrix multiplication. The polynomial dependence on p crucially allows our algorithms to extend naturally to efficient algorithms for ℓ_∞ regression, via approximation of ℓ_∞ by $\ell_{\mathcal{O}(\log n)}$. Of practical interest, we also develop a new subsampling algorithm for ℓ_p regression for arbitrary matrices, which is simpler than previous approaches for $p \ge 4$.

****

CrossBeam: Learning to Search in Bottom-Up Program Synthesis

Kensen Shi,Hanjun Dai,Kevin Ellis,Charles Sutton

Many approaches to program synthesis perform a search within an enormous space of programs to find one that satisfies a given specification. Prior works have used neural models to guide combinatorial search algorithms, but such approaches still explore a huge portion of the search space and quickly become intractable as the size of the desired program increases. To tame the search space blowup, we propose training a neural model to learn a hands-on search policy for bottom-up synthesis, instead of relying on a combinatorial search algorithm. Our approach, called CrossBeam, uses the neural model to choose how to combine previously-explored programs into new programs, taking into account the search history and partial program executions. Motivated by work in structured prediction on learning to search, CrossBeam is trained on-policy using data extracted from its own bottom-up searches on training tasks. We evaluate CrossBeam in two very different domains, string manipulation and logic programming. We observe that CrossBeam learns to search efficiently, exploring much smaller portions of the program space compared to the state-of-the-art.

****

PEARL: Data Synthesis via Private Embeddings and Adversarial Reconstruction Lear

ning
Seng Pei Liew,Tsubasa Takahashi,Michihiko Ueno
We propose a new framework of synthesizing data using deep generative models in a differentially private manner.
Within our framework, sensitive data are sanitized with rigorous privacy guarantees in a one-shot fashion, such that training deep generative models is possible without re-using the original data.
Hence, no extra privacy costs or model constraints are incurred, in contrast to popular gradient sanitization approaches, which, among other issues, cause degradation in privacy guarantees as the training iteration increases.
We demonstrate a realization of our framework by making use of the characteristic function and an adversarial re-weighting objective, which are of independent interest as well.
Our proposal has theoretical guarantees of performance, and empirical evaluations on multiple datasets show that our approach outperforms other methods at reasonable levels of privacy.
****
I-PGD-AT: Efficient Adversarial Training via Imitating Iterative PGD Attack
Xiaosen Wang,Bhavya Kailkhura,Krishnaram Kenthapadi,Bo Li
Adversarial training has been widely used in various machine learning paradigms to improve the robustness; while it would increase the training cost due to the perturbation optimization process. To improve the efficiency, recent studies leverage Fast Gradient Sign Method with Random Start (FGSM-RS) for adversarial training. However, such methods would lead to relatively low robustness and catastrophic overfitting, which means the robustness against iterative attacks (e.g. Projected Gradient Descent (PGD)) would suddenly drop to 0%. Different approaches have been proposed to address this problem, while later studies show that catastrophic overfitting still remains. In this paper, motivated by the fact that expensive iterative adversarial training methods achieve high robustness without catastrophic overfitting, we aim to ask: Can we perform iterative adversarial training in an efficient way? To this end, we first analyze the difference of perturbation generated by FGSM-RS and PGD and find that PGD tends to craft diverse discrete values instead of ± 1 in FGSM-RS. Based on this observation, we propose an efficient single-step adversarial training method I-PGD-AT by adopting I-PGD attack for training, in which I-PGD imitates PGD virtually. Unlike FGSM that crafts the perturbation directly using the sign of gradient, I-PGD imitates the perturbation of PGD based on the magnitude of gradient. Extensive empirical evaluations on CIFAR-10 and Tiny ImageNet demonstrate that our I-PGD-AT can improve the robustness compared with the baselines and significantly delay catastrophic overfitting. Moreover, we explore and discuss the factors that affect catastrophic overfitting. Finally, to demonstrate the generality of I-PGD-AT, we integrate it into PGD adversarial training and show that it can even further improve the robustness.
****
Why so pessimistic? Estimating uncertainties for offline RL through ensembles, and why their independence matters.
Seyed Kamyar Seyed Ghasemipour,Shixiang Shane Gu,Ofir Nachum
In order to achieve strong performance in offline reinforcement learning (RL), it is necessary to act conservatively with respect to confident lower-bounds on anticipated values of actions. Thus, a valuable approach would be to obtain high quality uncertainty estimates on action values. In current supervised learning literature, state-of-the-art approaches to uncertainty estimation and calibration rely on ensembling methods. In this work, we aim to transfer the success of ensembles from supervised learning to the setting of batch RL. We propose, MSG, a model-free dynamic programming based offline RL method that trains an ensemble of independent Q-functions, and updates a policy to act conservatively with respect to the uncertainties derived from the ensemble. Theoretically, by referring to the literature on infinite-width neural networks, we demonstrate the crucial dependence of the quality of uncertainty on the manner in which ensembling is performed, a phenomenon that arises due to the dynamic programming nature of RL and

overlooked by existing offline RL methods. Our theoretical predictions are corroborated by pedagogical examples on toy MDPs, as well as empirical comparisons in benchmark continuous control domains. In the more challenging domains of the D4RL offline RL benchmark, MSG significantly surpasses highly well-tuned state-of-the-art methods in batch RL. Motivated by the success of MSG, we investigate whether efficient approximations to ensembles can be as effective. We demonstrate that while efficient variants outperform current state-of-the-art, they do not match MSG with deep ensembles. We hope our work engenders increased focus into deep network uncertainty estimation techniques directed for reinforcement learning.

****

A Large Batch Optimizer Reality Check: Traditional, Generic Optimizers Suffice Across Batch Sizes

Zachary Nado,Justin Gilmer,Christopher J Shallue,Rohan Anil,George Edward Dahl

Recently the LARS and LAMB optimizers have been proposed for training neural networks faster using large batch sizes. LARS and LAMB add layer-wise normalization to the update rules of Heavy-ball momentum and Adam, respectively, and have become popular in prominent benchmarks and deep learning libraries. However, without fair comparisons to standard optimizers, it remains an open question whether LARS and LAMB have any benefit over traditional, generic algorithms. In this work we demonstrate that standard optimization algorithms such as Nesterov momentum and Adam can match or exceed the results of LARS and LAMB at large batch sizes. Our results establish new, stronger baselines for future comparisons at these batch sizes and shed light on the difficulties of comparing optimizers for neural network training more generally.

****

Improving greedy core-set configurations for active learning with uncertainty-scaled distances

Yuchen Li,Frank Rudzicz

We scale perceived distances of the core-set algorithm by a factor of uncertainty and search for low-confidence configurations, finding significant improvements in sample efficiency across CIFAR10/100 and SVHN image classification, especially in larger acquisition sizes. We show the necessity of our modifications and explain how the improvement is due to a probabilistic quadratic speed-up in the convergence of core-set loss, under assumptions about the relationship of model uncertainty and misclassification.

****

On the Relationship between Heterophily and Robustness of Graph Neural Networks

Jiong Zhu,Junchen Jin,Donald Loveland,Michael T Schaub,Danai Koutra

Empirical studies on the robustness of graph neural networks (GNNs) have suggested a relation between the vulnerabilities of GNNs to adversarial attacks and the increased presence of heterophily in perturbed graphs (where edges tend to connect nodes with dissimilar features and labels). In this work, we formalize the relation between heterophily and robustness, bridging two topics previously investigated by separate lines of research. We theoretically and empirically show that for graphs exhibiting homophily (low heterophily), impactful structural attacks always lead to increased levels of heterophily, while for graph with heterophily the change in the homophily level depends on the node degrees. By leveraging these insights, we deduce that a design principle identified to significantly improve predictive performance under heterophily—separate aggregators for ego- and neighbor-embeddings—can also inherently offer increased robustness to GNNs. Our extensive empirical analysis shows that GNNs adopting this design alone can achieve significantly improved empirical and certifiable robustness compared to the best-performing unvaccinated model. Furthermore, models with this design can be readily combined with explicit defense mechanisms to yield improved robustness with up to 18.33% increase in performance under attacks compared to the best-performing vaccinated model.

****

Word Sense Induction with Knowledge Distillation from BERT

Anik Saha,Alex Gittens,Bulent Yener

Pre-trained contextual language models are ubiquitously employed for language understanding tasks, but are unsuitable for resource-constrained systems. Noncontextual word embeddings are an efficient alternative in these settings. Such methods typically use one vector to encode multiple different meanings of a word, and incur errors due to polysemy. This paper proposes a two-stage method to distill multiple word senses from a pre-trained language model (BERT) by using attention over the senses of a word in a context and transferring this sense information to fit multi-sense embeddings in a skip-gram-like framework. We demonstrate an effective approach to training the sense disambiguation mechanism in our model with a distribution over word senses extracted from the output layer embeddings of BERT. Experiments on the contextual word similarity and sense induction tasks show that this method is superior to or competitive with state-of-the-art multi-sense embeddings on multiple benchmark data sets, and experiments with an embedding-based topic model (ETM) demonstrates the benefits of using this multi-sense embedding in a downstream application.

****

Divisive Feature Normalization Improves Image Recognition Performance in AlexNet

Michelle Miller,SueYeon Chung,Kenneth D. Miller

Local divisive normalization provides a phenomenological description of many nonlinear response properties of neurons across visual cortical areas. To gain insight into the utility of this operation, we studied the effects on AlexNet of a local divisive normalization between features, with learned parameters. Developing features were arranged in a line topology, with the influence between features determined by an exponential function of the distance between them. We compared an AlexNet model with no normalization or with canonical normalizations (Batch, Group, Layer) to the same models with divisive normalization added. Divisive normalization always improved performance for models with batch or group or no normalization, generally by 1-2 percentage points, on both the CIFAR-100 and ImageNet databases. To gain insight into mechanisms underlying the improved performance, we examined several aspects of network representations. In the early layers both canonical and divisive normalizations reduced manifold capacities and increased average dimension of the individual categorical manifolds. In later layers the capacity was higher and manifold dimension lower for models roughly in order of their performance improvement. Examining the sparsity of activations across a given layer, divisive normalization layers increased sparsity, while the canonical normalization layers decreased it. Nonetheless, in the final layer, the sparseness of activity increased in the order of no normalization, divisive, com- bined, and canonical. We also investigated how the receptive fields (RFs) in the first convolutional layer (where RFs are most interpretable) change with normalization. Divisive normalization enhanced RF Fourier power at low wavelengths, while divisive+canonical enhanced power at mid (batch, group) or low (layer) wavelengths, compared to canonical alone or no normalization. In conclusion, divisive normalization enhances image recognition performance, most strongly when combined with canonical normalization, and in doing so it reduces manifold capacity and sparsity in early layers while increasing them in final layers, and increases low- or mid-wavelength power in the first-layer receptive fields.

****

Learning Surface Parameterization for Document Image Unwarping

Sagnik Das,Ke Ma,Zhixin Shu,Dimitris Samaras

In this paper, we present a novel approach to learn texture mapping for a 3D surface and apply it to document image unwarping. We propose an efficient method to learn surface parameterization by learning a continuous bijective mapping between 3D surface positions and 2D texture-space coordinates. Our surface parameterization network can be conveniently plugged into a differentiable rendering pipeline and trained using multi-view images and rendering loss. Recent work on differentiable rendering techniques for implicit surfaces has shown high-quality 3D scene reconstruction and view synthesis results. However, these methods typically learn the appearance color as a function of the surface points and lack explicit surface parameterization. Thus they do not allow texture map extraction or tex ture editing. By introducing explicit surface parameterization and learning with a recent differentiable renderer for implicit surfaces, we demonstrate state-of-the-art document-unwarping via texture extraction. We show that our approach can reconstruct high-frequency textures for arbitrary document shapes in both synthetic and real scenarios. We also demonstrate the usefulness of our system by applying it to document texture editing.

****

EqR: Equivariant Representations for Data-Efficient Reinforcement Learning

Arnab Kumar Mondal,Vineet Jain,Kaleem Siddiqi,Siamak Ravanbakhsh

We study different notions of equivariance as an inductive bias in Reinforcement Learning (RL) and propose new mechanisms for recovering representations that are equivariant to both an agent’s action, and symmetry transformations of the state-action pairs. Whereas prior work on exploiting symmetries in deep RL can only incorporate predefined linear transformations, our approach allows for non-linear symmetry transformations of state-action pairs to be learned from the data itself. This is achieved through an equivariant Lie algebraic parameterization of state and action encodings, equivariant latent transition models, and the use of symmetry-based losses. We demonstrate the advantages of our learned equivariant representations for Atari games, in a data-efficient setting limited to 100k steps of interactions with the environment. Our method, which we call Equivariant representations for RL (EqR), outperforms many previous methods in a similar setting by achieving a median human-normalized score of 0.418, and surpassing human-level performance on 8 out of the 26 games.

****

Explore and Control with Adversarial Surprise

Arnaud Fickinger,Natasha Jaques,Samyak Parajuli,Michael Chang,Nicholas Rhinehart,Glen Berseth,Stuart Russell,Sergey Levine

Unsupervised reinforcement learning (RL) studies how to leverage environment statistics to learn useful behaviors without the cost of reward engineering. However, a central challenge in unsupervised RL is to extract behaviors that meaningfully affect the world and cover the range of possible outcomes, without getting distracted by inherently unpredictable, uncontrollable, and stochastic elements in the environment. To this end, we propose an unsupervised RL method designed for high-dimensional, stochastic environments based on an adversarial game between two policies (which we call Explore and Control) controlling a single body and competing over the amount of observation entropy the agent experiences. The Explore agent seeks out states that maximally surprise the Control agent, which in turn aims to minimize surprise, and thereby manipulate the environment to return to familiar and predictable states. The competition between these two policies drives them to seek out increasingly surprising parts of the environment while learning to gain mastery over them. We show formally that the resulting algorithm maximizes coverage of the underlying state in block MDPs with stochastic observations, providing theoretical backing to our hypothesis that this procedure avoids uncontrollable and stochastic distractions. Our experiments further demonstrate that Adversarial Surprise leads to the emergence of complex and meaningful skills, and outperforms state-of-the-art unsupervised reinforcement learning methods in terms of both exploration and zero-shot transfer to downstream tasks.

****

Evaluating Distributional Distortion in Neural Language Modeling

Benjamin LeBrun,Alessandro Sordoni,Timothy J. O'Donnell

A fundamental characteristic of natural language is the high rate at which speakers produce novel expressions. Because of this novelty, a heavy-tail of rare events accounts for a significant amount of the total probability mass of distributions in language (Baayen, 2001). Standard language modeling metrics such as perplexity quantify the performance of language models (LM) in aggregate. As a result, we have relatively little understanding of whether neural LMs accurately estimate the probability of sequences in this heavy-tail of rare events. To address this gap, we develop a controlled evaluation scheme which uses generative models trained on natural data as artificial languages from which we can exactly compute sequence probabilities. Training LMs on generations from these artificial la nguages, we compare the sequence-level probability estimates given by LMs to the true probabilities in the target language. Our experiments reveal that LSTM and Transformer language models (i) systematically underestimate the probability of sequences drawn from the target language, and (ii) do so more severely for less-probable sequences. Investigating where this probability mass went, (iii) we find that LMs tend to overestimate the probability of ill formed (perturbed) sequences. In addition, we find that this underestimation behaviour (iv) is weakened, but not eliminated by greater amounts of training data, and (v) is exacerbated for target distributions with lower entropy.

****

Logical Activation Functions: Logit-space equivalents of Boolean Operators

Scott C Lowe,Robert Earle,Jason d'Eon,Thomas Trappenberg,Sageev Oore

Neuronal representations within artificial neural networks are commonly understood as logits, representing the log-odds score of presence (versus absence) of features within the stimulus. Under this interpretation, we can derive the probability $P(x_0 \cap x_1)$ that a pair of independent features are both present in the stimulus from their logits. By converting the resulting probability back into a logit, we obtain a logit-space equivalent of the AND operation. However, since this function involves taking multiple exponents and logarithms, it is not well suited to be directly used within neural networks. We thus constructed an efficient approximation named AND_AIL (the AND operator Approximate for Independent Logits) utilizing only comparison and addition operations, which can be deployed as an activation function in neural networks. Like MaxOut, AND_AIL is a generalization of ReLU to two-dimensions. Additionally, we constructed efficient approximations of the logit-space equivalents to the OR and XNOR operators. We deployed these new activation functions, both in isolation and in conjunction, and demonstrated their effectiveness on a variety of tasks including image classification, transfer learning, abstract reasoning, and compositional zero-shot learning.

****

Continual Learning Using Task Conditional Neural Networks

Honglin Li,Frieder Ganz,David J. Sharp,Payam M. Barnaghi

Conventional deep learning models have limited capacity in learning multiple tasks sequentially. The issue of forgetting the previously learned tasks in continual learning is known as catastrophic forgetting or interference. When the input data or the goal of learning changes, a continual model will learn and adapt to the new status. However, the model will not remember or recognise any revisits to the previous states. This causes performance reduction and re-training curves in dealing with periodic or irregularly reoccurring changes in the data or goals. Dynamic approaches, which assign new neuron resources to the upcoming tasks, are introduced to address this issue. However, most of the dynamic methods need task information about the upcoming tasks during the inference phase to activate the corresponding neurons. To address this issue, we introduce Task Conditional Neural Network which allows the model to identify the task information automatically. The proposed model can continually learn and embed new tasks into the model without losing the information about previously learned tasks. We evaluate the proposed model combined with the mixture of experts approach on the MNIST and CIFAR100 datasets and show how it significantly improves the continual learning process without requiring task information in advance.

****

MaGNET: Uniform Sampling from Deep Generative Network Manifolds Without Retraining

Ahmed Imtiaz Humayun,Randall Balestriero,Richard Baraniuk

Deep Generative Networks (DGNs) are extensively employed in Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and their variants to approximate the data manifold, and data distribution on that manifold. However, training samples are often obtained based on preferences, costs, or convenience producing artifacts in the empirical data distribution e.g. the large fraction of smiling faces in the CelebA dataset or the large fraction of dark-haired individuals in FFHQ). {\em These inconsistencies will be reproduced when sampling from the trained DGN, which has far-reaching potential implications for fairness, data augmentation, anomaly detection, domain adaptation, and beyond.} In response, we develop a differential geometry based sampler -coined MaGNET- that, given any trained DGN, produces samples that are uniformly distributed on the learned manifold. We prove theoretically and empirically that our technique produces a uniform distribution on the manifold regardless of the training set distribution. We perform a range of experiments on various datasets and DGNs. One of them considers the state-of-the-art StyleGAN2 trained on FFHQ dataset, where uniform sampling via MaGNET increases distribution precision \& recall by 4.12\% \& 3.01\% and decreases gender bias by 41.2\%, without requiring labels or retraining.

****

Sampling with Mirrored Stein Operators

Jiaxin Shi,Chang Liu,Lester Mackey

We introduce a new family of particle evolution samplers suitable for constrained domains and non-Euclidean geometries. Stein Variational Mirror Descent and Mirrored Stein Variational Gradient Descent minimize the Kullback-Leibler (KL) divergence to constrained target distributions by evolving particles in a dual space defined by a mirror map. Stein Variational Natural Gradient exploits non-Euclidean geometry to more efficiently minimize the KL divergence to unconstrained targets. We derive these samplers from a new class of mirrored Stein operators and adaptive kernels developed in this work. We demonstrate that these new samplers yield accurate approximations to distributions on the simplex, deliver valid confidence intervals in post-selection inference, and converge more rapidly than prior methods in large-scale unconstrained posterior inference. Finally, we establish the convergence of our new procedures under verifiable conditions on the target distribution.

****

Planning in Stochastic Environments with a Learned Model

Ioannis Antonoglou,Julian Schrittwieser,Sherjil Ozair,Thomas K Hubert,David Silver

Model-based reinforcement learning has proven highly successful. However, learning a model in isolation from its use during planning is problematic in complex environments. To date, the most effective techniques have instead combined value-equivalent model learning with powerful tree-search methods. This approach is exemplified by MuZero, which has achieved state-of-the-art performance in a wide range of domains, from board games to visually rich environments, with discrete and continuous action spaces, in online and offline settings. However, previous instantiations of this approach were limited to the use of deterministic models. This limits their performance in environments that are inherently stochastic, partially observed, or so large and complex that they appear stochastic to a finite agent. In this paper we extend this approach to learn and plan with stochastic models. Specifically, we introduce a new algorithm, Stochastic MuZero, that learns a stochastic model incorporating afterstates, and uses this model to perform a stochastic tree search. Stochastic MuZero matched or exceeded the state of the art in a set of canonical single and multi-agent environments, including 2048 and backgammon, while maintaining the same performance as standard MuZero in the game of Go.

****

Neural Contextual Bandits with Deep Representation and Shallow Exploration

Pan Xu,Zheng Wen,Handong Zhao,Quanquan Gu

We study neural contextual bandits, a general class of contextual bandits, where each context-action pair is associated with a raw feature vector, but the specific reward generating function is unknown. We propose a novel learning algorithm that transforms the raw feature vector using the last hidden layer of a deep ReLU neural network (deep representation learning), and uses an upper confidence bound (UCB) approach to explore in the last linear layer (shallow exploration). We prove that under standard assumptions, our proposed algorithm achieves $\tilde{O}(\sqrt{T})$ finite-time regret, where T is the learning time horizon. Compared with existing neural contextual bandit algorithms, our approach is computationally much more efficient since it only needs to explore in the last layer of t he deep neural network.

****

PI3NN: Out-of-distribution-aware Prediction Intervals from Three Neural Networks
Siyan Liu,Pei Zhang,Dan Lu,Guannan Zhang

We propose a novel prediction interval (PI) method for uncertainty quantification, which addresses three major issues with the state-of-the-art PI methods. First, existing PI methods require retraining of neural networks (NNs) for every given confidence level and suffer from the crossing issue in calculating multiple PIs. Second, they usually rely on customized loss functions with extra sensitive hyperparameters for which fine tuning is required to achieve a well-calibrated PI. Third, they usually underestimate uncertainties of out-of-distribution (OOD) samples leading to over-confident PIs. Our PI3NN method calculates PIs from linear combinations of three NNs, each of which is independently trained using the standard mean squared error loss. The coefficients of the linear combinations are computed using root-finding algorithms to ensure tight PIs for a given confidence level. We theoretically prove that PI3NN can calculate PIs for a series of confidence levels without retraining NNs and it completely avoids the crossing issue. Additionally, PI3NN does not introduce any unusual hyperparameters resulting in a stable performance. Furthermore, we address OOD identification challenge by introducing an initialization scheme which provides reasonably larger PIs of the OOD samples than those of the in-distribution samples. Benchmark and real-world experiments show that our method outperforms several state-of-the-art approaches with respect to predictive uncertainty quality, robustness, and OOD samples identification.

****

Non-Transferable Learning: A New Approach for Model Ownership Verification and Applicability Authorization
Lixu Wang,Shichao Xu,Ruiqi Xu,Xiao Wang,Qi Zhu

As Artificial Intelligence as a Service gains popularity, protecting well-trained models as intellectual property is becoming increasingly important. There are two common types of protection methods: ownership verification and usage authorization. In this paper, we propose Non-Transferable Learning (NTL), a novel approach that captures the exclusive data representation in the learned model and restricts the model generalization ability to certain domains. This approach provides effective solutions to both model verification and authorization. Specifically: 1) For ownership verification, watermarking techniques are commonly used but are often vulnerable to sophisticated watermark removal methods. By comparison, our NTL-based ownership verification provides robust resistance to state-of-the-art watermark removal methods, as shown in extensive experiments with 6 removal approaches over the digits, CIFAR10 & STL10, and VisDA datasets. 2) For usage authorization, prior solutions focus on authorizing specific users to access the model, but authorized users can still apply the model to any data without restriction. Our NTL-based authorization approach instead provides data-centric protection, which we call applicability authorization, by significantly degrading the performance of the model on unauthorized data. Its effectiveness is also shown through experiments on aforementioned datasets.

****

Discriminative Similarity for Data Clustering
Yingzhen Yang,Ping Li

Similarity-based clustering methods separate data into clusters according to the pairwise similarity between the data, and the pairwise similarity is crucial for their performance. In this paper, we propose {\em Clustering by Discriminative Similarity (CDS)}, a novel method which learns discriminative similarity for data clustering. CDS learns an unsupervised similarity-based classifier from each data partition, and searches for the optimal partition of the data by minimizing the generalization error of the learnt classifiers associated with the data partitions. By generalization analysis via Rademacher complexity, the generalization error bound for the unsupervised similarity-based classifier is expressed as the sum of discriminative similarity between the data from different classes. It is proved that the derived discriminative similarity can also be induced by the integrated squared error bound for kernel density classification. In order to evaluate the performance of the proposed discriminative similarity, we propose a new clustering method using a kernel as the similarity function, CDS via unsupervised kernel classification (CDSK), with its effectiveness demonstrated by experimental results.

****

It Takes Four to Tango: Multiagent Self Play for Automatic Curriculum Generation

Yuqing Du,Pieter Abbeel,Aditya Grover

We are interested in training general-purpose reinforcement learning agents that can solve a wide variety of goals. Training such agents efficiently requires automatic generation of a goal curriculum. This is challenging as it requires (a) exploring goals of increasing difficulty, while ensuring that the agent (b) is exposed to a diverse set of goals in a sample efficient manner and (c) does not catastrophically forget previously solved goals. We propose Curriculum Self Play (CuSP), an automated goal generation framework that seeks to satisfy these desiderata by virtue of a multi-player game with 4 agents. We extend the asymmetric curricula learning in PAIRED (Dennis et al., 2020) to a symmetrized game that carefully balances cooperation and competition between two off-policy student learners and two regret-maximizing teachers. CuSP additionally introduces entropic goal coverage and accounts for the non-stationary nature of the students, allowing us to automatically induce a curriculum that balances progressive exploration with anti-catastrophic exploitation. We demonstrate that our method succeeds at generating an effective curricula of goals for a range of control tasks, outperforming other methods at zero-shot test-time generalization to novel out-of-distribution goals.

****

A Boosting Approach to Reinforcement Learning

Nataly Brukhim,Elad Hazan,Karan Singh

We study efficient algorithms for reinforcement learning in Markov decision processes, whose complexity is independent of the number of states. This formulation succinctly captures large scale problems, but is also known to be computationally hard in its general form.

Previous approaches attempt to circumvent the computational hardness by assuming structure in either transition function or the value function, or by relaxing the solution guarantee to a local optimality condition.

We consider the methodology of boosting, borrowed from supervised learning, for converting weak learners into an effective policy. The notion of weak learning we study is that of sampled-based approximate optimization of linear functions over policies. Under this assumption of weak learnability, we give an efficient algorithm that is capable of improving the accuracy of such weak learning methods iteratively. We prove sample complexity and running time bounds on our method, that are polynomial in the natural parameters of the problem: approximation guarantee, discount factor, distribution mismatch and number of actions. In particular, our bound does not explicitly depend on the number of states.

A technical difficulty in applying previous boosting results, is that the value function over policy space is not convex. We show how to use a non-convex variant of the Frank-Wolfe method, coupled with recent advances in gradient boosting that allow incorporating a weak learner with multiplicative approximation guarantee, to overcome the non-convexity and attain global optimality guarantees.

****

Universal Joint Approximation of Manifolds and Densities by Simple Injective Flows

Michael Anthony Puthawala,Matti Lassas,Ivan Dokmani■,Maarten V. de Hoop

We analyze neural networks composed of bijective flows and injective expansive elements. We find that such networks universally approximate a large class of manifolds simultaneously with densities supported on them. Among others, our results apply to the well-known coupling and autoregressive flows. We build on the work of Teshima et al. 2020 on bijective flows and study injective architectures pr oposed in Brehmer et al. 2020 and Kothari et al. 2021. Our results leverage a new theoretical device called the \emph{embedding gap}, which measures how far one continuous manifold is from embedding another. We relate the embedding gap to a relaxation of universally we call the \emph{manifold embedding property}, capturing the geometric part of universality. Our proof also establishes that optimality of a network can be established ``in reverse,'' resolving a conjecture made in Brehmer et al. 2020 and opening the door for simple layer-wise training schemes. Finally, we show that the studied networks admit an exact layer-wise projection result, Bayesian uncertainty quantification, and black-box recovery of network weights.

****

GrASP: Gradient-Based Affordance Selection for Planning

Vivek Veeriah,Zeyu Zheng,Richard Lewis,Satinder Singh

Planning with a learned model is arguably a key component of intelligence. There are several challenges in realizing such a component in large-scale reinforcement learning (RL) problems. One such challenge is dealing effectively with continuous action spaces when using tree-search planning (e.g., it is not feasible to consider every action even at just the root node of the tree). In this paper we present a method for \emph{selecting} affordances useful for planning---for learning which small number of actions/options from a continuous space of actions/options to consider in the tree-expansion process during planning. We consider affordances that are goal-and-state-conditional mappings to actions/options as well as unconditional affordances that simply select actions/options available in all states. Our selection method is gradient based: we compute gradients through the planning procedure to update the parameters of the function that represents affordances. Our empirical work shows that it is feasible to learn to select both primitive-action and option affordances, and that simultaneously learning to select affordances and planning with a learned value-equivalent model can outperform model-free RL.

****

Metric Learning on Temporal Graphs via Few-Shot Examples

Dongqi Fu,Liri Fang,Ross Maciejewski,Vetle I Torvik,Jingrui He

Graph metric learning methods aim to learn the distance metric over graphs such that similar graphs are closer and dissimilar graphs are farther apart. This is of critical importance in many graph classification applications such as drug discovery and epidemics categorization. In many real-world applications, the graphs are typically evolving over time; labeling graph data is usually expensive and also requires background knowledge. However, state-of-the-art graph metric learning techniques consider the input graph as static, and largely ignore the intrinsic dynamics of temporal graphs; Furthermore, most of these techniques require abundant labeled examples for training in the representation learning process. To address the two aforementioned problems, we wish to learn a distance metric only over fewer temporal graphs, which metric could not only help accurately categorize seen temporal graphs but also be adapted smoothly to unseen temporal graphs. In this paper, we first propose the streaming-snapshot model to describe temporal graphs on different time scales. Then we propose the MetaTag framework: 1) to learn the metric over a limited number of streaming-snapshot modeled temporal graphs, 2) and adapt the learned metric to unseen temporal graphs via a few examples. Finally, we demonstrate the performance of MetaTag in comparison with state-of-the-art algorithms for temporal graph classification problems.

****

Scalable multimodal variational autoencoders with surrogate joint posterior

Masahiro Suzuki,Yutaka Matsuo

To obtain a joint representation from multimodal data in variational autoencoders (VAEs), it is important to infer the representation from arbitrary subsets of modalities after learning. A scalable way to achieve this is to aggregate the inferences of each modality as experts. A state-of-the-art approach to learning this aggregation of experts is to encourage all modalities to be reconstructed and cross-generated from arbitrary subsets. However, this learning may be insufficient if cross-generation is difficult. Furthermore, to evaluate its objective fu nction, exponential generation paths concerning the number of modalities are required. To alleviate these problems, we propose to explicitly minimize the divergence between inferences from arbitrary subsets and the surrogate joint posterior that approximates the true joint posterior. We also proposed using a gradient origin network, a deep generative model that learns inferences without using an inference network, thereby reducing the need for additional parameters by introducing the surrogate posterior. We demonstrate that our method performs better than existing scalable multimodal VAEs in inference and generation.

****

Semi-supervised Offline Reinforcement Learning with Pre-trained Decision Transformers

Catherine Cang,Kourosh Hakhamaneshi,Ryan Rudes,Igor Mordatch,Aravind Rajeswaran,Pieter Abbeel,Michael Laskin

Pre-training deep neural network models using large unlabelled datasets followed by fine-tuning them on small task-specific datasets has emerged as a dominant paradigm in natural language processing (NLP) and computer vision (CV). Despite the widespread success, such a paradigm has remained atypical in reinforcement learning (RL).

In this paper, we investigate how we can leverage large reward-free (i.e. task-agnostic) offline datasets of prior interactions to pre-train agents that can then be fine-tuned using a small reward-annotated dataset. To this end, we present Pre-trained Decision Transformer (PDT), a simple yet powerful algorithm for semi-supervised Offline RL. By masking reward tokens during pre-training, the transformer learns to autoregressivley predict actions based on previous state and action context and effectively extracts behaviors present in the dataset. During fine-tuning, rewards are un-masked and the agent learns the set of skills that should be invoked for the desired behavior as per the reward function. We demonstrate the efficacy of this simple and flexible approach on tasks from the D4RL benchmark with limited reward annotations.

****

Increase and Conquer: Training Graph Neural Networks on Growing Graphs

Juan Cervino,Luana Ruiz,Alejandro Ribeiro

Graph neural networks (GNNs) use graph convolutions to exploit network invariances and learn meaningful features from network data. However, on large-scale graphs convolutions incur in high computational cost, leading to scalability limitations. Leveraging the graphon --- the limit object of a graph --- in this paper we consider the problem of learning a graphon neural network (WNN) --- the limit object of a GNN --- by training GNNs on graphs sampled Bernoulli from the graphon. Under smoothness conditions, we show that: (i) the expected distance between the learning steps on the GNN and on the WNN decreases asymptotically with the size of the graph, and (ii) when training on a sequence of growing graphs, gradient descent follows the learning direction of the WNN. Inspired by these results, we propose a novel algorithm to learn GNNs on large-scale graphs that, starting from a moderate number of nodes, successively increases the size of the graph during training. This algorithm is benchmarked on both a recommendation system and a decentralized control problem where it is shown to retain comparable performance, to its large-scale counterpart, at a reduced computational cost.

****

CROP: Certifying Robust Policies for Reinforcement Learning through Functional Smoothing

Fan Wu,Linyi Li,Zijian Huang,Yevgeniy Vorobeychik,Ding Zhao,Bo Li

As reinforcement learning (RL) has achieved great success and been even adopted in safety-critical domains such as autonomous vehicles, a range of empirical studies have been conducted to improve its robustness against adversarial attacks. However, how to certify its robustness with theoretical guarantees still remains challenging. In this paper, we present the first unified framework CROP (Certifying Robust Policies for RL) to provide robustness certification on both action and reward levels. In particular, we propose two robustness certification criteria: robustness of per-state actions and lower bound of cumulative rewards. We then d evelop a local smoothing algorithm for policies derived from Q-functions to guarantee the robustness of actions taken along the trajectory; we also develop a global smoothing algorithm for certifying the lower bound of a finite-horizon cumulative reward, as well as a novel local smoothing algorithm to perform adaptive search in order to obtain tighter reward certification. Empirically, we apply CROP to evaluate several existing empirically robust RL algorithms, including adversarial training and different robust regularization, in four environments (two representative Atari games, Highway, and CartPole). Furthermore, by evaluating these algorithms against adversarial attacks, we demonstrate that our certifications are often tight. All experiment results are available at website https://crop-leaderboard.github.io.

****

Multi-Trigger-Key: Towards Multi-Task Privacy-Preserving In Deep Learning

Ren Wang,Zhe Xu,Alfred Hero

Deep learning-based Multi-Task Classification (MTC) is widely used in applications like facial attribute and healthcare that warrant strong privacy guarantees. In this work, we aim to protect sensitive information in the inference phase of MTC and propose a novel Multi-Trigger-Key (MTK) framework to achieve the privacy-preserving objective. MTK associates each secured task in the multi-task dataset with a specifically designed trigger-key. The true information can be revealed by adding the trigger-key if the user is authorized. We obtain such an MTK model by training it with a newly generated training set. To address the information leakage malaise resulting from correlations among different tasks, we generalize the training process by incorporating an MTK decoupling process with a controllable trade-off between the protective efficacy and the model performance. Theoretical guarantees and experimental results demonstrate the effectiveness of the privacy protection without appreciable hindering on the model performance.

****

Towards General Robustness to Bad Training Data

Tianhao Wang,Yi Zeng,Ming Jin,Ruoxi Jia

In this paper, we focus on the problem of identifying bad training data when the underlying cause is unknown in advance. Our key insight is that regardless of how bad data are generated, they tend to contribute little to training a model with good prediction performance or more generally, to some utility function of the data analyst. We formulate the problem of good/bad data selection as utility optimization. We propose a theoretical framework for evaluating the worst-case performance of data selection heuristics. Remarkably, our results show that the popular heuristic based on the Shapley value may choose the worst data subset in certain practical scenarios, which sheds lights on its large performance variation observed empirically in the past work. We then develop an algorithmic framework, DataSifter, to detect a variety of and even unknown data issues---a step towards general robustness to bad training data. DataSifter is guided by the theoretically optimal solution to data selection and is made practical by the data utility learning technique. Our evaluation shows that DataSifter achieves and most often significantly improves the state-of-the-art performance over a wide range of tasks, including backdoor, poison, noisy/mislabel data detection, data summarization, and data debiasing.

****

Multiresolution Equivariant Graph Variational Autoencoder

Truong Son Hy,Risi Kondor

In this paper, we propose Multiresolution Equivariant Graph Variational Autoencoders (MGVAE), the first hierarchical generative model to learn and generate graphs in a multiresolution and equivariant manner. At each resolution level, MGVAE employs higher order message passing to encode the graph while learning to partition it into mutually exclusive clusters and coarsening into a lower resolution that eventually creates a hierarchy of latent distributions. MGVAE then constructs a hierarchical generative model to variationally decode into a hierarchy of coarsened graphs. Importantly, our proposed framework is end-to-end permutation equivariant with respect to node ordering. MGVAE achieves competitive results with several generative tasks including general graph generation, molecular generat ion, unsupervised molecular representation learning to predict molecular properties, link prediction on citation graphs, and graph-based image generation.

****

Unifying Top-down and Bottom-up for Recurrent Visual Attention

GANG CHEN

The idea of using the recurrent neural network for visual attention has gained popularity in computer vision community. Although the recurrent visual attention model (RAM) leverages the glimpses with more large patch size to increasing its scope, it may result in high variance and instability. For example, we need the Gaussian policy with high variance to explore object of interests in a large image, which may cause randomized search and unstable learning. In this paper, we propose to unify the top-down and bottom-up attention together for recurrent visual attention. Our model exploits the image pyramids and Q-learning to select regions of interests in the top-down attention mechanism, which in turn to guide the policy search in the bottom-up approach. In addition, we add another two constraints over the bottom-up recurrent neural networks for better exploration. We train our model in an end-to-end reinforcement learning framework, and evaluate our method on visual classification tasks. The experimental results outperform convolutional neural networks (CNNs) baseline and the bottom-up recurrent models with visual attention.

****

Specialized Transformers: Faster, Smaller and more Accurate NLP Models

Amrit Nagarajan,Sanchari Sen,Jacob R. Stevens,Anand Raghunathan

Transformers have greatly advanced the state-of-the-art in Natural Language Processing (NLP) in recent years, but are especially demanding in terms of their computation and storage requirements. Transformers are first pre-trained on a large dataset, and subsequently fine-tuned for different downstream tasks. We observe that this design process leads to models that are not only over-parameterized for downstream tasks, but also contain elements that adversely impact accuracy of the downstream tasks.

We propose a Specialization framework to create optimized transformer models for a given downstream task. Our framework systematically uses accuracy-driven pruning, i.e., it identifies and prunes parts of the pre-trained Transformer that hinder performance on the downstream task. We also replace the dense soft-attention in selected layers with sparse hard-attention to help the model focus on the relevant parts of the input. In effect, our framework leads to models that are not only faster and smaller, but also more accurate. The large number of parameters contained in Transformers presents a challenge in the form of a large pruning design space. Further, the traditional iterative prune-retrain approach is not applicable to Transformers, since the fine-tuning data is often very small and re-training quickly leads to overfitting. To address these challenges, we propose a hierarchical, re-training-free pruning method with model- and task- specific heuristics. Our experiments on GLUE and SQUAD show that Specialized models are consistently more accurate (by up to 4.5\%), while also being up to 2.5$\times$ faster and up to 3.2$\times$ smaller than the conventional fine-tuned models. In addition, we demonstrate that Specialization can be combined with previous efforts such as distillation or quantization to achieve further benefits.

For example, Specialized Q8BERT and DistilBERT models exceed the performance of BERT-Base, while being up to 3.7$\times$ faster and up to 12.1$\times$ smaller.

****

Geometric Random Walk Graph Neural Networks via Implicit Layers

Giannis Nikolentzos,Michalis Vazirgiannis

Graph neural networks have recently attracted a lot of attention and have been applied with great success to several important graph problems. The Random Walk Graph Neural Network model was recently proposed as a more intuitive alternative to the well-studied family of message passing neural networks. This model compares each input graph against a set of latent ``hidden graphs'' using a kernel that counts common random walks up to some length. In this paper, we propose a new architecture, called Geometric Random Walk Graph Neural Network (GRWNN), that generalizes the above model such that it can count common walks of infinite length in two graphs. The proposed model retains the transparency of Random Walk Graph Neural Networks since its first layer also consists of a number of trainable ``hidden graphs'' which are compared against the input graphs using the geometric random walk kernel. To compute the kernel, we employ a fixed-point iteration approach involving implicitly defined operations. Then, we capitalize on implicit differentiation to derive an efficient training scheme which requires only constant memory, regardless of the number of fixed-point iterations. The employed random walk kernel is differentiable, and therefore, the proposed model is end-to-end trainable. Experiments on standard graph classification datasets demonstrate the effectiveness of the proposed approach in comparison with state-of-the-art methods.

****

Curriculum Discovery through an Encompassing Curriculum Learning Framework
Mohamed Elgaar,Hadi Amiri

We describe a curriculum learning framework capable of discovering optimal curricula in addition to performing standard curriculum learning. We show that this framework encompasses existing curriculum learning approaches such as difficulty-based data sub-sampling, data pruning, and loss re-weighting. We employ the proposed framework to address the following key questions in curriculum learning research: (a) What is the best curriculum to train a given model on a given dataset? (b) What are the characteristics of optimal curricula for different datasets and different difficulty scoring functions? We show that our framework outperforms competing state-of-the-art curriculum learning approaches in natural language inference and other text classification tasks. In addition, exhaustive experiments illustrate the generalizability of the discovered curricula across the three datasets and two difficulty scoring functions.

****

Pretraining for Language Conditioned Imitation with Transformers
Aaron L Putterman,Kevin Lu,Igor Mordatch,Pieter Abbeel

We study reinforcement learning (RL) agents which can utilize language inputs. To investigate this, we propose a new multimodal benchmark -- Text-Conditioned Frostbite -- in which an agent must complete tasks specified by text instructions in the Atari Frostbite environment. We curate and release a dataset of 5M text-labelled transitions for training and to encourage further research in this direction. On this benchmark, we evaluate Text Decision Transformer (TDT), a transformer directly operating on text, state, and action tokens, and find it improves upon other baseline architectures. Furthermore, we evaluate the effect of pretraining, finding unsupervised pretraining can yield improved results in low-data settings.

****

The weighted mean trick – optimization strategies for robustness
Valeriu Balaban,Paul Bogdan

We prove that minimizing a weighted mean results in optimizing the higher-order moments of the loss distribution such as the variance, skewness, and kurtosis. By optimizing the higher-order moments, one can tighten the upper bound of the loss mean deviating from the true expectation and improve the robustness against outliers. Such types of optimization problems often lead to non-convex objectives, therefore, we explore the extent to which the proposed weighted mean trick preserves convexity, albeit at times at a decrease in efficiency. Experimental results show that the weighted mean trick exhibits similar performance with other specialized robust loss functions when training on noisy datasets while providing a stronger theoretical background. The proposed weighted mean trick is a simple yet powerful optimization framework that is easy to integrate into existing works.

****

Detecting Worst-case Corruptions via Loss Landscape Curvature in Deep Reinforcement Learning
Ezgi Korkmaz,Jonah Brown-Cohen

The non-robustness of neural network policies to adversarial examples poses a challenge for deep reinforcement learning. One natural approach to mitigate the impact of adversarial examples is to develop methods to detect when a given input is adversarial. In this work we introduce a novel approach for detecting adversarial examples that is computationally efficient, is agnostic to the method used to generate adversarial examples, and theoretically well-motivated. Our method is based on a measure of the local curvature of the neural network policy, which we show differs between adversarial and clean examples. We empirically demonstrate the effectiveness of our method in the Atari environment against a large set of state-of-the-art algorithms for generating adversarial examples. Furthermore, we exhibit the effectiveness of our detection algorithm with the presence of multiple strong detection-aware adversaries.
****

Neural Link Prediction with Walk Pooling

Liming Pan,Cheng Shi,Ivan Dokmani■

Graph neural networks achieve high accuracy in link prediction by jointly leveraging graph topology and node attributes. Topology, however, is represented indirectly; state-of-the-art methods based on subgraph classification label nodes with distance to the target link, so that, although topological information is present, it is tempered by pooling. This makes it challenging to leverage features like loops and motifs associated with network formation mechanisms. We propose a link prediction algorithm based on a new pooling scheme called WalkPool. WalkPool combines the expressivity of topological heuristics with the feature-learning ability of neural networks. It summarizes a putative link by random walk probabilities of adjacent paths. Instead of extracting transition probabilities from the original graph, it computes the transition matrix of a ``predictive'' latent graph by applying attention to learned features; this may be interpreted as feature-sensitive topology fingerprinting. WalkPool can leverage unsupervised node features or be combined with GNNs and trained end-to-end. It outperforms state-of-the-art methods on all common link prediction benchmarks, both homophilic and heterophilic, with and without node attributes. Applying WalkPool to a set of unsupervised GNNs significantly improves prediction accuracy, suggesting that it may be used as a general-purpose graph pooling scheme.
****

Graph Piece: Efficiently Generating High-Quality Molecular Graphs with Substructures

Xiangzhe Kong,Zhixing Tan,Yang Liu

Molecular graph generation is a fundamental but challenging task in various applications such as drug discovery and material science, which requires generating valid molecules with desired properties. Auto-regressive models, which usually construct graphs following sequential actions of adding nodes and edges at the atom-level, have made rapid progress in recent years. However, these atom-level models ignore high-frequency subgraphs that not only capture the regularities of atomic combination in molecules but also are often related to desired chemical properties. In this paper, we propose a method to automatically discover such common substructures, which we call graph pieces, from given molecular graphs. Based on graph pieces, we leverage a variational autoencoder to generate molecules in two phases: piece-level graph generation followed by bond completion. Experiments show that our graph piece variational autoencoder achieves better performance over state-of-the-art baselines on property optimization and constrained property optimization tasks with higher computational efficiency.
****

Influence-Based Reinforcement Learning for Intrinsically-Motivated Agents

Ammar Fayad,Majd Ibrahim

Discovering successful coordinated behaviors is a central challenge in Multi-Agent Reinforcement Learning (MARL) since it requires exploring a joint action space that grows exponentially with the number of agents. In this paper, we propose a mechanism for achieving sufficient exploration and coordination in a team of agents. Specifically, agents are rewarded for contributing to a more diversified team behavior by employing proper intrinsic motivation functions. To learn meani ngful coordination protocols, we structure agents’ interactions by introducing a novel framework, where at each timestep, an agent simulates counterfactual rollouts of its policy and, through a sequence of computations, assesses the gap between other agents’ current behaviors and their targets. Actions that minimize the gap are considered highly influential and are rewarded. We evaluate our approach on a set of challenging tasks with sparse rewards and partial observability that require learning complex cooperative strategies under a proper exploration scheme, such as the StarCraft Multi-Agent Challenge. Our methods show significantly improved performances over different baselines across all tasks.

****

Learning Efficient and Robust Ordinary Differential Equations via Diffeomorphisms

Weiming Zhi,Tin Lai,Lionel Ott,Edwin V Bonilla,Fabio Ramos

Advances in differentiable numerical integrators have enabled the use of gradient descent techniques to learn ordinary differential equations (ODEs), where a flexible function approximator (often a neural network) is used to estimate the system dynamics, given as a time derivative. However, these integrators can be unsatisfactorily slow and unstable when learning systems of ODEs from long sequences. We propose to learn an ODE of interest from data by viewing its dynamics as a vector field related to another \emph{base} vector field via a diffeomorphism (i.e., a differentiable bijection). By learning both the diffeomorphism and the dynamics of the base ODE, we provide an avenue to offload some of the complexity in modelling the dynamics directly on to learning the diffeomorphism. Consequently, by restricting the base ODE to be amenable to integration, we can speed up and improve the robustness of integrating trajectories from the learned system. We demonstrate the efficacy of our method in training and evaluating benchmark ODE systems, as well as within continuous-depth neural networks models. We show that our approach attains speed-ups of up to two orders of magnitude when integrating learned ODEs.

****

AutoOED: Automated Optimal Experimental Design Platform with Data- and Time-Efficient Multi-Objective Optimization

Yunsheng Tian,Mina Konakovic Lukovic,Michael Foshey,Timothy Erps,Beichen Li,Wojciech Matusik

We present AutoOED, an Automated Optimal Experimental Design platform powered by machine learning to accelerate discovering solutions with optimal objective trade-offs. To solve expensive multi-objective problems in a data-efficient manner, we implement popular multi-objective Bayesian optimization (MOBO) algorithms with state-of-the-art performance in a modular framework. To further accelerate the optimization in a time-efficient manner, we propose a novel strategy called Believer-Penalizer (BP), which allows batch experiments to be accelerated asynchronously without affecting performance. AutoOED serves as a testbed for machine learning researchers to quickly develop and evaluate their own MOBO algorithms. We also provide a graphical user interface (GUI) for users with little or no experience with coding, machine learning, or optimization to visualize and guide the experiment design intuitively. Finally, we demonstrate that AutoOED can control and guide real-world hardware experiments in a fully automated way without human intervention.

****

A Frequency Perspective of Adversarial Robustness

Shishira Maiya,Max Ehrlich,Vatsal Agarwal,Ser-Nam Lim,Tom Goldstein,Abhinav Shrivastava

Adversarial examples pose a unique challenge for deep learning systems. Despite recent advances in both attacks and defenses, there is still a lack of clarity and consensus in the community about the true nature and underlying properties of adversarial examples. A deep understanding of these examples can provide new insights towards the development of more effective attacks and defenses. Driven by the common misconception that adversarial examples are high-frequency noise, we present a frequency-based understanding of adversarial examples, supported by theoretical and empirical findings. Our analysis shows that adversarial examples are neither in high-frequency nor in low-frequency components, but are simply dataset dependent. Particularly, we highlight the glaring disparities between models trained on CIFAR-10 and ImageNet-derived datasets. Utilizing this framework, we analyze many intriguing properties of training robust models with frequency constraints, and propose a frequency-based explanation for the commonly observed accuracy vs. robustness trade-off.

****

Sharp Attention for Sequence to Sequence Learning
Pei Zhang,Hua Liu

Attention mechanism has been widely applied to tasks that output some sequence from an input image. Its success comes from the ability to align relevant parts of the encoded image with the target output. However, most of the existing methods fail to build clear alignment because the aligned parts are unable to well represent the target. In this paper we seek clear alignment in attention mechanism through a \emph{sharpener} module. Since it deliberately locates the target in an image region and refines representation to be target-specific, the alignment and interpretability of attention can be significantly improved. Experiments on synthetic handwritten digit as well as real-world scene text recognition datasets show that our approach outperforms the mainstream ones such as soft and hard attention.

****

On the Convergence of Certified Robust Training with Interval Bound Propagation
Yihan Wang,Zhouxing Shi,Quanquan Gu,Cho-Jui Hsieh

Interval Bound Propagation (IBP) is so far the base of state-of-the-art methods for training neural networks with certifiable robustness guarantees when potential adversarial perturbations present, while the convergence of IBP training remains unknown in existing literature. In this paper, we present a theoretical analysis on the convergence of IBP training. With an overparameterized assumption, we analyze the convergence of IBP robust training. We show that when using IBP training to train a randomly initialized two-layer ReLU neural network with logistic loss, gradient descent can linearly converge to zero robust training error with a high probability if we have sufficiently small perturbation radius and large network width.

****

Pretraining Text Encoders with Adversarial Mixture of Training Signal Generators
Yu Meng,Chenyan Xiong,Payal Bajaj,saurabh tiwary,Paul N. Bennett,Jiawei Han,Xia Song

We present a new framework AMOS that pretrains text encoders with an Adversarial learning curriculum via a Mixture Of Signals from multiple auxiliary generators. Following ELECTRA-style pretraining, the main encoder is trained as a discriminator to detect replaced tokens generated by auxiliary masked language models (MLMs). Different from ELECTRA which trains one MLM as the generator, we jointly train multiple MLMs of different sizes to provide training signals at various levels of difficulty. To push the discriminator to learn better with challenging replaced tokens, we learn mixture weights over the auxiliary MLMs' outputs to maximize the discriminator loss by backpropagating the gradient from the discriminator via Gumbel-Softmax. For better pretraining efficiency, we propose a way to assemble multiple MLMs into one unified auxiliary model. AMOS outperforms ELECTRA and recent state-of-the-art pretrained models by about 1 point on the GLUE benchmark for BERT base-sized models.

****

Spectral Bias in Practice: the Role of Function Frequency in Generalization
Sara Fridovich-Keil,Raphael Gontijo-Lopes,Rebecca Roelofs

Despite their ability to represent highly expressive functions, deep learning models trained with SGD seem to find simple, constrained 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 only been observed in theoretical models and simplified experiments. In this work, we propose methodologies for measuring spectral bias in modern image classification networks. We find that these networks indeed exhibit spectral bias, and that networks that generalize well strike a balance between having enough complexity (i.e. high frequencies) to fit the data while being simple enough to avoid overfitting. For example, we experimentally show that larger models learn high frequencies faster than smaller ones, but many forms of regularization, both explicit and implicit, amplify spectral bias and delay the learning of high frequencies. 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. Our work enables measuring and ultimately controlling the spectral behavior of neural networks used for image classification, and is a step towards understanding why deep models generalize well.

****

Patches Are All You Need?

Asher Trockman,J Zico Kolter

Although convolutional networks have been the dominant architecture for vision tasks for many years, recent experiments have shown that Transformer-based models, most notably the Vision Transformer (ViT), may exceed their performance in some settings. However, due to the quadratic runtime of the self-attention layers in Transformers, ViTs require the use of patch embeddings, which group together small regions of the image into single input features, in order to be applied to larger image sizes. This raises a question: Is the performance of ViTs due to the inherently-more-powerful Transformer architecture, or is it at least partly due to using patches as the input representation? In this paper, we present some evidence for the latter: specifically, we propose the ConvMixer, an extremely simple model that is similar in spirit to the ViT and the even-more-basic MLP-Mixer in that it operates directly on patches as input, separates the mixing of spatial and channel dimensions, and maintains equal size and resolution throughout the network. In contrast, however, the ConvMixer uses only standard convolutions to achieve the mixing steps. Despite its simplicity, we show that the ConvMixer outperforms the ViT, MLP-Mixer, and some of their variants for similar parameter counts and data set sizes, in addition to outperforming classical vision models such as the ResNet. Our code is available at https://github.com/tmp-iclr/convmixer.

****

Maximum Likelihood Estimation for Multimodal Learning with Missing Modality

Fei Ma,Xiangxiang Xu,Shao-Lun Huang,Lin Zhang

Multimodal learning has achieved great successes in many scenarios. Compared with unimodal learning, it can effectively combine the information from different modalities to improve the performance of learning tasks. In reality, the multimodal data may have missing modalities due to various reasons, such as sensor failure and data transmission error. In previous works, the information of the modality-missing data has not been well exploited. To address this problem, we propose an efficient approach based on maximum likelihood estimation to incorporate the knowledge in the modality-missing data. Specifically, we design a likelihood function to characterize the conditional distributions of the modality-complete data and the modality-missing data, which is theoretically optimal. Moreover, we develop a generalized form of the softmax function to effectively implement maximum likelihood estimation in an end-to-end manner. Such training strategy guarantees the computability of our algorithm capably. Finally, we conduct a series of experiments on real-world multimodal datasets. Our results demonstrate the effectiveness of the proposed approach, even when 95% of the training data has missing modality.

****

Target Propagation via Regularized Inversion

Vincent Roulet,Zaid Harchaoui

Target Propagation (TP) algorithms compute targets instead of gradients along neural networks and propagate them backward in a way that is similar yet different than gradient back-propagation (BP). The idea was first presented as a perturbative alternative to back-propagation that may achieve greater accuracy in gradient evaluation when training multi-layer neural networks (LeCun et al., 1989). Ho

wever, TP has remained more of a template algorithm with many variations than a well-identified algorithm. Revisiting insights of LeCun et al., (1989) and more recently of Lee et al. (2015), we present a simple version of target propagation based on regularized inversion of network layers, easily implementable in a differentiable programming framework. We compare its computational complexity to the one of BP and delineate the regimes in which TP can be attractive compared to BP. We show how our TP can be used to train recurrent neural networks with long sequences on various sequence modeling problems. The experimental results underscore the importance of regularization in TP in practice.
****

Neural Structured Prediction for Inductive Node Classification
Meng Qu,Huiyu Cai,Jian Tang
This paper studies node classification in the inductive setting, i.e., aiming to learn a model on labeled training graphs and generalize it to infer node labels on unlabeled test graphs. This problem has been extensively studied with graph neural networks (GNNs) by learning effective node representations, as well as traditional structured prediction methods for modeling the structured output of node labels, e.g., conditional random fields (CRFs). In this paper, we present a new approach called the Structured Proxy Network (SPN), which combines the advantages of both worlds. SPN defines flexible potential functions of CRFs with GNNs. However, learning such a model is nontrivial as it involves optimizing a maximin game with high-cost inference. Inspired by the underlying connection between joint and marginal distributions defined by Markov networks, we propose to solve an approximate version of the optimization problem as a proxy, which yields a near-optimal solution, making learning more efficient. Extensive experiments on two settings show that our approach outperforms many competitive baselines.
****

Towards Training Billion Parameter Graph Neural Networks for Atomic Simulations
Anuroop Sriram,Abhishek Das,Brandon M Wood,Siddharth Goyal,C. Lawrence Zitnick
Recent progress in Graph Neural Networks (GNNs) for modeling atomic simulations has the potential to revolutionize catalyst discovery, which is a key step in making progress towards the energy breakthroughs needed to combat climate change. However, the GNNs that have proven most effective for this task are memory intensive as they model higher-order interactions in the graphs such as those between triplets or quadruplets of atoms, making it challenging to scale these models. In this paper, we introduce Graph Parallelism, a method to distribute input graphs across multiple GPUs, enabling us to train very large GNNs with hundreds of millions or billions of parameters. We empirically evaluate our method by scaling up the recently proposed DimeNet++ and GemNet models by over an order of magnitude in the number of parameters. On the large-scale Open Catalyst 2020 (OC20) dataset, these graph-parallelized models lead to relative improvements of 1) 15% on the force MAE metric on the S2EF task and 2) 21% on the AFbT metric on the IS2RS task, establishing new state-of-the-art results.
****

Low-Cost Algorithmic Recourse for Users With Uncertain Cost Functions
Prateek Yadav,Peter Hase,Mohit Bansal
The problem of identifying algorithmic recourse for people affected by machine learning model decisions has received much attention recently. Existing approaches for recourse generation obtain solutions using properties like diversity, proximity, sparsity, and validity. Yet, these objectives are only heuristics for what we truly care about, which is whether a user is satisfied with the recourses offered to them. Some recent works try to model user-incurred cost, which is more directly linked to user satisfaction. But they assume a single global cost function that is shared across all users. This is an unrealistic assumption when users have dissimilar preferences about their willingness to act upon a feature and different costs associated with changing that feature. In this work, we formalize the notion of user-specific cost functions and introduce a new method for identifying actionable recourses for users. By default, we assume that users' cost functions are hidden from the recourse method, though our framework allows users to partially or completely specify their preferences or cost function. We propo

se an objective function, Expected Minimum Cost (EMC), based on two key ideas: (1) when presenting a set of options to a user, it is vital that there is at least one low-cost solution the user could adopt; (2) when we do not know the user's true cost function, we can approximately optimize for user satisfaction by first sampling plausible cost functions, then finding a set that achieves a good cost for the user in expectation. We optimize EMC with a novel discrete optimization algorithm, Cost-Optimized Local Search (COLS), which is guaranteed to improve the recourse set quality over iterations. Experimental evaluation on popular real-world datasets with simulated user costs demonstrates that our method satisfies up to 25.89 percentage points more users compared to strong baseline methods.

Using standard fairness metrics, we also show that our method can provide more fair solutions across demographic groups than comparable methods, and we verify that our method is robust to misspecification of the cost function distribution.

****

L-SR1 Adaptive Regularization by Cubics for Deep Learning

Aditya Ranganath,Mukesh Singhal,Roummel Marcia

Stochastic gradient descent and other first-order variants, such as Adam and AdaGrad, are commonly used in the field of deep learning due to their computational efficiency and low-storage memory requirements. However, these methods do not exploit curvature information. Consequently, iterates can converge to saddle points and poor local minima. To avoid these points, directions of negative curvature can be utilized, which requires computing the second-derivative matrix. In Deep Neural Networks (DNNs), the number of variables (n) can be of the order of tens of millions, making the Hessian impractical to store ($\mathcal{O}(n^2)$) and to invert ($\mathcal{O}(n^3)$). Alternatively, quasi-Newton methods compute Hessian approximations that do not have the same computational requirements. Quasi-Newton methods re-use previously computed iterates and gradients to compute a low-rank structured update. The most widely used quasi-Newton update is the L-BFGS, which guarantees a positive semi-definite Hessian approximation, making it suitable in a line search setting. However, the loss function in DNNs are non-convex, where the Hessian is potentially non-positive definite. In this paper, we propose using a Limited-Memory Symmetric Rank-1 quasi-Newton approach which allows for indefinite Hessian approximations, enabling directions of negative curvature to be exploited. Furthermore, we use a modified Adaptive Regularized Cubics approach, which generates a sequence of cubic subproblems that have closed-form solutions. We investigate the performance of our proposed method on autoencoders and feed-forward neural network models and compare our approach to state-of-the-art first-order adaptive stochastic methods as well as L-BFGS.

****

Offline Reinforcement Learning with Resource Constrained Online Deployment

Jayanth Reddy Regatti,Aniket Anand Deshmukh,Young Hun Jung,Frank Cheng,Abhishek Gupta,Urun Dogan

Offline reinforcement learning is used to train policies in scenarios where real-time access to the environment is expensive or impossible.

As a natural consequence of these harsh conditions, an agent may lack the resources to fully observe the online environment before taking an action. We dub this situation the resource-constrained setting. This leads to situations where the offline dataset (available for training) can contain fully processed features (using powerful language models, image models, complex sensors, etc.) which are not available when actions are actually taken online.

This disconnect leads to an interesting and unexplored problem in offline RL: Is it possible to use a richly processed offline dataset to train a policy which has access to fewer features in the online environment?

In this work, we introduce and formalize this novel resource-constrained problem setting. We highlight the performance gap between policies trained using the full offline dataset and policies trained using limited features.

We address this performance gap with a policy transfer algorithm which first trains a teacher agent using the offline dataset where features are fully available, and then transfers this knowledge to a student agent that only uses the resource-constrained features. To better capture the challenge of this setting, we pro pose a data collection procedure: Resource Constrained-Datasets for RL (RC-D4RL) . We evaluate our transfer algorithm on RC-D4RL and the popular D4RL benchmarks and observe consistent improvement over the baseline (TD3+BC without transfer).

****

Efficient Certification for Probabilistic Robustness

Victor Rong,Alexandre Megretski,Luca Daniel,Tsui-Wei Weng

Recent developments on the robustness of neural networks have primarily emphasized the notion of worst-case adversarial robustness in both verification and robust training. However, often looser constraints are needed and some margin of error is allowed. We instead consider the task of probabilistic robustness, which assumes the input follows a known probabilistic distribution and seeks to bound the probability of a given network failing against the input. We focus on developing an efficient robustness verification algorithm by extending a bound-propagation-based approach. Our proposed algorithm improves upon the robustness certificate of this algorithm by up to $8\times$ while with no additional computational cost. In addition, we perform a case study on incorporating the probabilistic robustness verification during training for the first time.

****

RotoGrad: Gradient Homogenization in Multitask Learning

Adrián Javaloy,Isabel Valera

Multitask learning is being increasingly adopted in applications domains like computer vision and reinforcement learning. However, optimally exploiting its advantages remains a major challenge due to the effect of negative transfer. Previous works have tracked down this issue to the disparities in gradient magnitudes and directions across tasks, when optimizing the shared network parameters. While recent work has acknowledged that negative transfer is a two-fold problem, existing approaches fall short as they only focus on either homogenizing the gradient magnitude across tasks; or greedily change the gradient directions, overlooking future conflicts. In this work, we introduce RotoGrad, an algorithm that tackles negative transfer as a whole: it jointly homogenizes gradient magnitudes and directions, while ensuring training convergence. We show that RotoGrad outperforms competing methods in complex problems, including multi-label classification in CelebA and computer vision tasks in the NYUv2 dataset. A Pytorch implementation can be found in https://github.com/adrianjav/rotograd.

****

Factored World Models for Zero-Shot Generalization in Robotic Manipulation

Ondrej Biza,Thomas Kipf,David Klee,Robert Platt,Jan-Willem van de Meent,Lawson L.S. Wong

World models for environments with many objects face a combinatorial explosion of states: as the number of objects increases, the number of possible arrangements grows exponentially. In this paper, we learn to generalize over robotic pick-and-place tasks using object-factored world models, which combat the combinatorial explosion by ensuring that predictions are equivariant to permutations of objects. We build on one such model, C-SWM, which we extend to overcome the assumption that each action is associated with one object. To do so, we introduce an action attention module to determine which objects are likely to be affected by an action. The attention module is used in conjunction with a residual graph neural network block that receives action information at multiple levels. Based on RGB images and parameterized motion primitives, our model can accurately predict the dynamics of a robot building structures from blocks of various shapes. Our model generalizes over training structures built in different positions. Moreover crucially, the learned model can make predictions about tasks not represented in training data. That is, we demonstrate successful zero-shot generalization to novel tasks. For example, we measure only 2.4% absolute decrease in our action ranking metric in the case of a block assembly task.

****

Disentangling One Factor at a Time

Vaishnavi S Patil,Matthew S Evanusa,Joseph JaJa

With the overabundance of data for machines to process in the current state of machine learning, data discovery, organization, and interpretation of the data be comes a critical need. Specifically of need are unsupervised methods that do not require laborious labeling by human observers. One promising approach to this enedeavour is \textit{Disentanglement}, which aims at learning the underlying generative latent factors of the data. The factors should also be as human interpretable as possible for the purposes of data discovery. \textit{Unsupervised disentanglement} is a particularly difficult open subset of the problem, which asks the network to learn on its own the generative factors without any link to the true labels. This problem area is currently dominated by two approaches: Variational Autoencoder and Generative Adversarial Network approaches. While GANs have good performance, they suffer from difficulty in training and mode collapse, and while VAEs are stable to train, they do not perform as well as GANs in terms of interpretability. In current state of the art versions of these approaches, the networks require the user to specify the number of factors that we expect to find in the data. This limitation prevents "true" disentanglement, in the sense that learning how many factors is actually one of the tasks we wish the network to solve. In this work we propose a novel network for unsupervised disentanglement that combines the stable training of the VAE with the interpretability offered by GANs without the training instabilities. We aim to disentangle interpretable latent factors "one at a time", or OAT factor learning, making no prior assumptions about the number or distribution of factors, in a completely unsupervised manner. We demonstrate its quantitative and qualitative effectiveness by evaluating the latent representations learned on two benchmark datasets, DSprites and CelebA.

****

Learning Sample Reweighting for Adversarial Robustness

Chester Holtz,Tsui-Wei Weng,Gal Mishne

There has been great interest in enhancing the robustness of neural network classifiers to defend against adversarial perturbations through adversarial training, while balancing the trade-off between robust accuracy and standard accuracy. We propose a novel adversarial training framework that learns to reweight the loss associated with individual training samples based on a notion of class-conditioned margin, with the goal of improving robust generalization. Inspired by MAML-based approaches, we formulate weighted adversarial training as a bilevel optimization problem where the upper-level task corresponds to learning a robust classifier, and the lower-level task corresponds to learning a parametric function that maps from a sample's \textit{multi-class margin} to an importance weight. Extensive experiments demonstrate that our approach improves both clean and robust accuracy compared to related techniques and state-of-the-art baselines.

****

On Improving Adversarial Transferability of Vision Transformers

Muzammal Naseer,Kanchana Ranasinghe,Salman Khan,Fahad Khan,Fatih Porikli

Vision transformers (ViTs) process input images as sequences of patches via self-attention; a radically different architecture than convolutional neural networks (CNNs). This makes it interesting to study the adversarial feature space of ViT models and their transferability. In particular, we observe that adversarial patterns found via conventional adversarial attacks show very \emph{low} black-box transferability even for large ViT models. We show that this phenomenon is only due to the sub-optimal attack procedures that do not leverage the true representation potential of ViTs. A deep ViT is composed of multiple blocks, with a consistent architecture comprising of self-attention and feed-forward layers, where each block is capable of independently producing a class token. Formulating an attack using only the last class token (conventional approach) does not directly leverage the discriminative information stored in the earlier tokens, leading to poor adversarial transferability of ViTs.Using the compositional nature of ViT models, we enhance transferability of existing attacks by introducing two novel strategies specific to the architecture of ViT models. \emph{(i) Self-Ensemble:} We propose a method to find multiple discriminative pathways by dissecting a single ViT model into an ensemble of networks. This allows explicitly utilizing class-specific information at each ViT block. \emph{(ii) Token Refinement:} We then propose to refine the tokens to further enhance the discriminative capacity at each block of ViT.Our token refinement systematically combines the class tokens with structural information preserved within the patch tokens. An adversarial attack when applied to such refined tokens within the ensemble of classifiers found in a single vision transformer has significantly higher transferability and thereby brings out the true generalization potential of the ViT's adversarial space. Code: https://t.ly/hBbW.

****

On Predicting Generalization using GANs
Yi Zhang,Arushi Gupta,Nikunj Saunshi,Sanjeev Arora

Research on generalization bounds for deep networks seeks to give ways to predict test error using just the training dataset and the network parameters. While generalization bounds can give many insights about architecture design, training algorithms etc., what they do not currently do is yield good predictions for actual test error. A recently introduced Predicting Generalization in Deep Learning competition aims to encourage discovery of methods to better predict test error. The current paper investigates a simple idea: can test error be predicted using {\em synthetic data,} produced using a Generative Adversarial Network (GAN) that was trained on the same training dataset? Upon investigating several GAN models and architectures, we find that this turns out to be the case.

In fact, using GANs pre-trained on standard datasets, the test error can be predicted without requiring any additional hyper-parameter tuning. This result is surprising because GANs have well-known limitations (e.g. mode collapse) and are known to not learn the data distribution accurately. Yet the generated samples are good enough to substitute for test data. Several additional experiments are presented to explore reasons why GANs do well at this task. In addition to a new approach for predicting generalization, the counter-intuitive phenomena presented in our work may also call for a better understanding of GANs' strengths and limitations.

****

Convergence Analysis and Implicit Regularization of Feedback Alignment for Deep Linear Networks
Manuela Girotti,Ioannis Mitliagkas,Gauthier Gidel

We theoretically analyze the Feedback Alignment (FA) algorithm, an efficient alternative to backpropagation for training neural networks. We provide convergence guarantees with rates for deep linear networks for both continuous and discrete dynamics. Additionally, we study incremental learning phenomena for shallow linear networks. Interestingly, certain specific initializations imply that negligible components are learned {before} the principal ones, thus potentially negatively affecting the effectiveness of such a learning algorithm; a phenomenon we classify as implicit anti-regularization. We also provide initialization schemes where the components of the problem are approximately learned by decreasing order of importance, thus providing a form of implicit regularization.

****

Learning a metacognition for object detection
Marlene Berke,Mario Belledonne,Zhangir Azerbayev,Julian Jara-Ettinger

In contrast to object recognition models, humans do not blindly trust their perception when building representations of the world, instead recruiting metacognition to detect percepts that are unreliable or false, such as when we realize that we mistook one object for another. We propose METAGEN, an unsupervised model that enhances object recognition models through a metacognition. Given noisy output from an object-detection model, METAGEN learns a meta-representation of how its perceptual system works and uses it to infer the objects in the world responsible for the detections. METAGEN achieves this by conditioning its inference on basic principles of objects that even human infants understand (known as Spelke principles: object permanence, cohesion, and spatiotemporal continuity). We test METAGEN on a variety of state-of-the-art object detection neural networks. We find that METAGEN quickly learns an accurate metacognitive representation of the neural network, and that this improves detection accuracy by filling in objects that the detection model missed and removing hallucinated objects. This approach enables generalization to out-of-sample data and outperforms comparison models that lack a metacognition.

****

Logarithmic landscape and power-law escape rate of SGD

Takashi Mori,Liu Ziyin,Kangqiao Liu,Masahito Ueda

Stochastic gradient descent (SGD) undergoes complicated multiplicative noise for the mean-square loss. We use this property of the SGD noise to derive a stochastic differential equation (SDE) with simpler additive noise by performing a random time change. In the SDE, the loss gradient is replaced by the logarithmized loss gradient. By using this formalism, we obtain the escape rate formula from a local minimum, which is determined not by the loss barrier height $\Delta L=L(\theta^s)-L(\theta^*)$ between a minimum θ^* and a saddle θ^s but by the logarithmized loss barrier height $\Delta\log L=\log[L(\theta^s)/L(\theta^*)]$. Our escape-rate formula strongly depends on the typical magnitude h^* and the number n of the outlier eigenvalues of the Hessian. This result explains an empirical fact that SGD prefers flat minima with low effective dimensions, which gives an insight into implicit biases of SGD.

****

WHY FLATNESS DOES AND DOES NOT CORRELATE WITH GENERALIZATION FOR DEEP NEURAL NETWORKS

Shuofeng Zhang,Isaac Reid,Guillermo Valle-Perez,Ard A. Louis

The intuition that local flatness of the loss landscape is correlated with better generalization for deep neural networks (DNNs) has been explored for decades, spawning many different flatness measures. Recently, this link with generalization has been called into question by a demonstration that many measures of flatness are vulnerable to parameter re-scaling which arbitrarily changes their value without changing neural network outputs. Here we show that, in addition, some popular variants of SGD such as Adam and Entropy-SGD, can also break the flatness-generalization correlation. As an alternative to flatness measures, we use a function based picture and propose using the log of Bayesian prior upon initialization, $\log P(f)$, as a predictor of the generalization when a DNN converges on function f after training to zero error.   The prior is directly proportional to the Bayesian posterior for functions that give zero error on a test set. For the case of image classification, we show that $\log P(f)$ is a significantly more robust predictor of generalization than flatness measures are. Whilst local flatness measures fail under parameter re-scaling, the prior/posterior, which is global quantity, remains invariant under re-scaling. Moreover, the correlation with generalization as a function of data complexity remains good for different variants of SGD.

****

Limitations of Active Learning With Deep Transformer Language Models

Mike D'Arcy,Doug Downey

Active Learning (AL) has the potential to reduce labeling cost when training natural language processing models, but its effectiveness with the large pretrained transformer language models that power today's NLP is uncertain. We present experiments showing that when applied to modern pretrained models, active learning offers inconsistent and often poor performance. As in prior work, we find that AL sometimes selects harmful "unlearnable" collective outliers, but we discover that some failures have a different explanation: the examples AL selects are informative but also increase training instability, reducing average performance. Our findings suggest that for some datasets this instability can be mitigated by training multiple models and selecting the best on a validation set, which we show impacts relative AL performance comparably to the outlier-pruning technique from prior work while also increasing absolute performance. Our experiments span three pretrained models, ten datasets, and four active learning approaches.

****

Understanding the Generalization of Adam in Learning Neural Networks with Proper Regularization

Difan Zou,Yuan Cao,Yuanzhi Li,Quanquan Gu

Adaptive gradient methods such as Adam have gained increasing popularity in deep learning optimization. However, it has been observed in many deep learning applications such as image classification, Adam can converge to a different solution with a worse test error compared to (stochastic) gradient descent, even with a fine-tuned regularization. In this paper, we provide a theoretical explanation for this phenomenon: we show that in the nonconvex setting of learning over-parameterized two-layer convolutional neural networks starting from the same random initialization, for a class of data distributions (inspired from image data), Adam and gradient descent (GD) can converge to different global solutions of the training objective with provably different generalization errors, even with weight decay regularization. In contrast, we show that if the training objective is convex, and the weight decay regularization is employed, any optimization algorithms including Adam and GD will converge to the same solution if the training is successful. This suggests that the generalization gap between Adam and SGD is fundamentally tied to the nonconvex landscape of deep learning optimization, which cannot be covered by the recent neural tangent kernel (NTK) based analysis.

****

Understanding and Leveraging Overparameterization in Recursive Value Estimation

Chenjun Xiao,Bo Dai,Jincheng Mei,Oscar A Ramirez,Ramki Gummadi,Chris Harris,Dale Schuurmans

The theory of function approximation in reinforcement learning (RL) typically considers low capacity representations that incur a tradeoff between approximation error, stability and generalization. Current deep architectures, however, operate in an overparameterized regime where approximation error is not necessarily a bottleneck. To better understand the utility of deep models in RL we present an analysis of recursive value estimation using \emph{overparameterized} linear representations that provides useful, transferable findings. First, we show that classical updates such as temporal difference (TD) learning or fitted-value-iteration (FVI) converge to \emph{different} fixed points than residual minimization (RM) in the overparameterized linear case. We then develop a unified interpretation of overparameterized linear value estimation as minimizing the Euclidean norm of the weights subject to alternative constraints. A practical consequence is that RM can be modified by a simple alteration of the backup targets to obtain the same fixed points as FVI and TD (when they converge), while universally ensuring stability. Further, we provide an analysis of the generalization error of these methods, demonstrating per iterate bounds on the value prediction error of FVI, and fixed point bounds for TD and RM.

Given this understanding, we then develop new algorithmic tools for improving recursive value estimation with deep models.

In particular, we extract two regularizers that penalize out-of-span top-layer weights and co-linearity in top-layer features respectively. Empirically we find that these regularizers dramatically improve the stability of TD and FVI, while allowing RM to match and even sometimes surpass their generalization performance with assured stability.

****

Exploring unfairness in Integrated Gradients based attribution methods

David Drakard,Rosanne Liu,Jason Yosinski

Numerous methods have attempted to explain and interpret predictions made by machine learning models in terms of their inputs. Known as “attribution methods” they notably include the Integrated Gradients method and its variants.These are based upon the theory of Shapley Values, a rigorous method of fair allocation according to mathematical axioms. Integrated Gradients has axioms derived from this heritage with the implication of a similar rigorous, intuitive notion of fairness. We explore the difference between Integrated Gradients and more direct expressions of Shapley Values in deep learning and find Integrated Gradients’ guarantees of fairness weaker; in certain conditions it can give wholly unrepresentative results. Integrated Gradients requires a choice of “baseline”, a hyperparameter that represents the ‘zero attribution’ case. Research has shown that baseline choice critically affects attribution quality, and increasingly effective baselines have been developed. Using purpose-designed scenarios we identify sources of inaccuracy both from specific baselines and inherent t

o the method itself, sensitive to input distribution and loss landscape. Failure modes are identified for baselines including Zero, Mean,Additive Gaussian Noise, and the state of the art Expected Gradients. We develop a new method, Integrated Certainty Gradients, that we show avoids the failures in these challenging scenarios. By augmenting the input space with “certainty”information, and training with random degradation of input features, the model learns to predict with varying amounts of incomplete information, supporting a zero-information case which becomes a natural baseline. We identify the axiomatic origin of unfairness in Integrated Gradients, which has been overlooked in past research.
****

Optimization and Adaptive Generalization of Three layer Neural Networks
Khashayar Gatmiry,Stefanie Jegelka,Jonathan Kelner

While there has been substantial recent work studying generalization of neural networks,

the ability of deep nets in automating the process of feature extraction still evades a thorough mathematical understanding.

As a step toward this goal, we analyze learning and generalization of a three-layer neural network with ReLU activations in a regime that goes beyond the linear approximation of the network, and is hence not captured by the common Neural Tangent Kernel. We show that despite nonconvexity of the empirical loss, a variant of SGD converges in polynomially many iterations to a good solution that generalizes. In particular, our generalization bounds are adaptive: they automatically optimize over a family of kernels that includes the Neural Tangent Kernel, to provide the tightest bound.
****

Conditional Expectation based Value Decomposition for Scalable On-Demand Ride Pooling
Avinandan Bose,Pradeep Varakantham

Owing to the benefits for customers (lower prices), drivers (higher revenues), aggregation companies (higher revenues) and the environment (fewer vehicles), on-demand ride pooling (e.g., Uber pool, Grab Share) has become quite popular. The significant computational complexity of matching vehicles to combinations of requests has meant that traditional ride pooling approaches are myopic in that they do not consider the impact of current matches on future value for vehicles/drivers.

Recently, Neural Approximate Dynamic Programming (NeurADP) has employed value decomposition with Approximate Dynamic Programming (ADP) to outperform leading approaches by considering the impact of an individual agent's (vehicle) chosen actions on the future value of that agent. However, in order to ensure scalability and facilitate city-scale ride pooling, NeurADP completely ignores the impact of other agents actions on individual agent/vehicle value. As demonstrated in our experimental results, ignoring the impact of other agents actions on individual value can have a significant impact on the overall performance when there is increased competition among vehicles for demand. Our key contribution is a novel mechanism based on computing conditional expectations through joint conditional probabilities for capturing dependencies on other agents actions without increasing the complexity of training or decision making. We show that our new approach, Conditional Expectation based Value Decomposition (CEVD) outperforms NeurADP by up to 9.76$\% $in terms of overall requests served, which is a significant improvement on a city wide benchmark taxi dataset.
****

Multi-scale Feature Learning Dynamics: Insights for Double Descent
Mohammad Pezeshki,Amartya Mitra,Yoshua Bengio,Guillaume Lajoie

A key challenge in building theoretical foundations for deep learning is the complex optimization dynamics of neural networks, resulting from the high-dimensional interactions between the large number of network parameters. Such non-trivial interactions lead to intriguing model behaviors such as the phenomenon of "double descent" of the generalization error. The more commonly studied aspect of this phenomenon corresponds to model-wise double descent where the test error exhi

bits a second descent with increasing model complexity, beyond the classical U-shaped error curve. In this work, we investigate the origins of the less studied epoch-wise double descent in which the test error undergoes two non-monotonous transitions, or descents as the training time increases. We study a linear teacher-student setup exhibiting epoch-wise double descent similar to that in deep neural networks. In this setting, we derive closed-form analytical expressions for the evolution of generalization error over training. We find that double descent can be attributed to distinct features being learned at different scales: as fast-learning features overfit, slower-learning features start to fit, resulting in a second descent in test error. We validate our findings through numerical experiments where our theory accurately predicts empirical findings and remains consistent with observations in deep neural networks.
****

Non-Parallel Text Style Transfer with Self-Parallel Supervision
Ruibo Liu,Chongyang Gao,Chenyan Jia,Guangxuan Xu,Soroush Vosoughi
The performance of existing text style transfer models is severely limited by the non-parallel datasets on which the models are trained. In non-parallel datasets, no direct mapping exists between sentences of the source and target style; the style transfer models thus only receive weak supervision of the target sentences during training, which often leads the model to discard too much style-independent information, or utterly fail to transfer the style.

In this work, we propose LaMer, a novel text style transfer framework based on large-scale language models. LaMer first mines the roughly parallel expressions in the non-parallel datasets with scene graphs, and then employs MLE training, followed by imitation learning refinement, to leverage the intrinsic parallelism within the data. On two benchmark tasks (sentiment & formality transfer) and a newly proposed challenging task (political stance transfer), our model achieves qualitative advances in transfer accuracy, content preservation, and fluency. Further empirical and human evaluations demonstrate that our model not only makes training more efficient, but also generates more readable and diverse expressions than previous models.
****

Can an Image Classifier Suffice For Action Recognition?
Quanfu Fan,Chun-Fu Chen,Rameswar Panda
We explore a new perspective on video understanding by casting the video recognition problem as an image recognition task. Our approach rearranges input video frames into super images, which allow for training an image classifier directly to fulfill the task of action recognition, in exactly the same way as image classification. With such a simple idea, we show that transformer-based image classifiers alone can suffice for action recognition. In particular, our approach demonstrates strong and promising performance against SOTA methods on several public datasets including Kinetics400, Moments In Time, Something-Something V2 (SSV2), Jester and Diving48. We also experiment with the prevalent ResNet image classifiers in computer vision to further validate our idea. The results on both Kinetics400 and SSV2 are comparable to some of the best-performed CNN approaches based on spatio-temporal modeling. Our source codes and models are available at \url{https://github.com/IBM/sifar-pytorch}.
****

Efficient and Modular Implicit Differentiation
Mathieu Blondel,Quentin Berthet,marco cuturi,Roy Frostig,Stephan Hoyer,Felipe Llinares-López,Fabian Pedregosa,Jean-Philippe Vert
Automatic differentiation (autodiff) has revolutionized machine learning. It allows expressing complex computations by composing elementary ones in creative ways and removes the tedious burden of computing their derivatives by hand. More recently, differentiation of optimization problem solutions has attracted a great deal of research, with applications as a layer in a neural network, and in bi-level optimization, including hyper-parameter optimization. However, the formulae for these derivatives often involves a tedious manual derivation and implementation. In this paper, we propose a unified, efficient and modular approach for im

plicit 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 to automatically differentiate the optimization problem. This way, our approach combines the benefits of implicit differentiation and autodiff. We show that seemingly simple principles allow to recover all recently proposed implicit differentiation methods and create new ones easily. We describe in details a JAX implementation of our framework and demonstrate the ease of differentiating through optimization problems thanks to it on four diverse tasks: hyperparameter optimization of multiclass SVMs, dataset distillation, task-driven dictionary learning and sensitivity analysis of molecular dynamics.

****

Avoiding Robust Misclassifications for Improved Robustness without Accuracy Loss

Yannick Merkli,Pavol Bielik,PETAR TSANKOV,Martin Vechev

While current methods for training robust deep learning models optimize robust accuracy, in practice, the resulting models are often both robust and inaccurate on numerous samples, providing a false sense of safety for those. Further, they significantly reduce natural accuracy, which hinders the adoption in practice. In this work, we address both of these challenges by extending prior works in three main directions. First, we propose a new training method that jointly maximizes robust accuracy and minimizes robust inaccuracy. Second, since the resulting models are trained to be robust only if they are accurate, we leverage robustness as a principled abstain mechanism. Finally, this abstain mechanism allows us to combine models in a compositional architecture that significantly boosts overall robustness without sacrificing accuracy. We demonstrate the effectiveness of our approach to both empirical and certified robustness on six recent state-of-the-art models and using several datasets. Our results show that our method effectively reduces robust and inaccurate samples by up to 97.28%. Further, it successfully enhanced the $\epsilon_\infty = 1/255$ robustness of a state-of-the-art model from 26% to 86% while only marginally reducing its natural accuracy from 97.8% to 97.6%.

****

Faster No-Regret Learning Dynamics for Extensive-Form Correlated Equilibrium

Ioannis Anagnostides,Gabriele Farina,Christian Kroer,Tuomas Sandholm

A recent emerging trend in the literature on learning in games has been concerned with providing accelerated learning dynamics for correlated and coarse correlated equilibria in normal-form games. Much less is known about the significantly more challenging setting of extensive-form games, which can capture sequential and simultaneous moves, as well as imperfect information. In this paper, we develop faster no-regret learning dynamics for \textit{extensive-form correlated equilibrium (EFCE)} in multiplayer general-sum imperfect-information extensive-form games. When all agents play T repetitions of the game according to the accelerated dynamics, the correlated distribution of play is an $O(T^{-3/4})$-approximate EFCE. This significantly improves over the best prior rate of $O(T^{-1/2})$. One of our conceptual contributions is to connect predictive (that is, optimistic) regret minimization with the framework of Φ-regret. One of our main technical contributions is to characterize the stability of certain fixed point strategies through a refined perturbation analysis of a structured Markov chain, which may be of independent interest.

Finally, experiments on standard benchmarks corroborate our findings.

****

On the Connection between Local Attention and Dynamic Depth-wise Convolution

Qi Han,Zejia Fan,Qi Dai,Lei Sun,Ming-Ming Cheng,Jiaying Liu,Jingdong Wang

Vision Transformer (ViT) attains state-of-the-art performance in visual recognition, and the variant, Local Vision Transformer, makes further improvements. The major component in Local Vision Transformer, local attention, performs the attention separately over small local windows. We rephrase local attention as a channel-wise locally-connected layer and analyze it from two network regularization manners, sparse connectivity and weight sharing, as well as dynamic weight computation. We point out that local attention resembles depth-wise convolution and it s dynamic variants in sparse connectivity: there is no connection across channels, and each position is connected to the positions within a small local window. The main differences lie in (i) weight sharing - depth-wise convolution shares connection weights (kernel weights) across spatial positions and attention shares the connection weights across channels, and (ii) dynamic weight computation manners - local attention is based on dot-products between pairwise positions in the local window, and dynamic convolution is based on linear projections conducted on the center representation or the globally pooled representation. The connection between local attention and dynamic depth-wise convolution is empirically verified by the ablation study about weight sharing and dynamic weight computation in Local Vision Transformer and (dynamic) depth-wise convolution. We empirically observe that the models based on depth-wise convolution and the dynamic variants with lower computation complexity perform on-par with or slightly better than Swin Transformer, an instance of Local Vision Transformer, for ImageNet classification, COCO object detection and ADE semantic segmentation. Code is available at https://github.com/Atten4Vis/DemystifyLocalViT.

****

Strength of Minibatch Noise in SGD

Liu Ziyin,Kangqiao Liu,Takashi Mori,Masahito Ueda

The noise in stochastic gradient descent (SGD), caused by minibatch sampling, is poorly understood despite its practical importance in deep learning. This work presents the first systematic study of the SGD noise and fluctuations close to a local minimum. We first analyze the SGD noise in linear regression in detail and then derive a general formula for approximating SGD noise in different types of minima. For application, our results (1) provide insight into the stability of training a neural network, (2) suggest that a large learning rate can help generalization by introducing an implicit regularization, (3) explain why the linear learning rate-batchsize scaling law fails at a large learning rate or at a small batchsize and (4) can provide an understanding of how discrete-time nature of SGD affects the recently discovered power-law phenomenon of SGD.

****

Learning more skills through optimistic exploration

DJ Strouse,Kate Baumli,David Warde-Farley,Volodymyr Mnih,Steven Stenberg Hansen

Unsupervised skill learning objectives (Eysenbach et al., 2019; Gregor et al., 2016) allow agents to learn rich repertoires of behavior in the absence of extrinsic rewards. They work by simultaneously training a policy to produce distinguishable latent-conditioned trajectories, and a discriminator to evaluate distinguishability by trying to infer latents from trajectories. The hope is for the agent to explore and master the environment by encouraging each skill (latent) to reliably reach different states. However, an inherent exploration problem lingers: when a novel state is actually encountered, the discriminator will necessarily not have seen enough training data to produce accurate and confident skill classifications, leading to low intrinsic reward for the agent and effective penalization of the sort of exploration needed to actually maximize the objective. To combat this inherent pessimism towards exploration, we derive an information gain auxiliary objective that involves training an ensemble of discriminators and rewarding the policy for their disagreement. Our objective directly estimates the epistemic uncertainty that comes from the discriminator not having seen enough training examples, thus providing an intrinsic reward more tailored to the true objective compared to pseudocount-based methods (Burda et al., 2019). We call this exploration bonus discriminator disagreement intrinsic reward, or DISDAIN. We demonstrate empirically that DISDAIN improves skill learning both in a tabular grid world (Four Rooms) and the 57 games of the Atari Suite (from pixels). Thus, we encourage researchers to treat pessimism with DISDAIN.

****

Iterative Sketching and its Application to Federated Learning

Zhao Song,Zheng Yu,Lichen Zhang

Johnson-Lindenstrauss lemma is one of the most valuable tools in machine learning, since it enables the reduction to the dimension of various learning problems. In this paper, we exploit the power of Fast-JL transform or so-called sketching technique and apply it to federated learning settings. Federated learning is an emerging learning scheme which allows multiple clients to train models without data exchange. Though most federated learning frameworks only require clients and the server to send gradient information over the network, they still face the challenges of communication efficiency and data privacy. We show that by iteratively applying independent sketches combined with additive noises, one can achieve the above two goals simultaneously. In our designed framework, each client only passes a sketched gradient to the server, and de-sketches the average-gradient information received from the server to synchronize. Such framework enjoys several benefits: 1). Better privacy, since we only exchange randomly sketched gradients with low-dimensional noises, which is more robust against emerging gradient attacks; 2). Lower communication cost per round, since our framework only communicates low-dimensional sketched gradients, which is particularly valuable in a small-bandwidth channel; 3). No extra overall communication cost. We provably show that the introduced randomness does not increase the overall communication at all.

****

ZeroSARAH: Efficient Nonconvex Finite-Sum Optimization with Zero Full Gradient Computations

Zhize Li,Slavomir Hanzely,Peter Richtárik

We propose ZeroSARAH -- a novel variant of the variance-reduced method SARAH (Nguyen et al., 2017) -- for minimizing the average of a large number of nonconvex functions $\frac{1}{n}\sum_{i=1}^{n}f_i(x)$. To the best of our knowledge, in this nonconvex finite-sum regime, all existing variance-reduced methods, including SARAH, SVRG, SAGA and their variants, need to compute the full gradient over all n data samples at the initial point x^0, and then periodically compute the full gradient once every few iterations (for SVRG, SARAH and their variants). Note that SVRG, SAGA and their variants typically achieve weaker convergence results than variants of SARAH: $n^{2/3}/\epsilon^2$ vs. $n^{1/2}/\epsilon^2$. Thus we focus on the variant of SARAH. The proposed ZeroSARAH and its distributed variant D-ZeroSARAH are the \emph{first} variance-reduced algorithms which \emph{do not require any full gradient computations}, not even for the initial point. Moreover, for both standard and distributed settings, we show that ZeroSARAH and D-ZeroSARAH obtain new state-of-the-art convergence results, which can improve the previous best-known result (given by e.g., SPIDER, SARAH, and PAGE) in certain regimes. Avoiding any full gradient computations (which are time-consuming steps) is important in many applications as the number of data samples n usually is very large. Especially in the distributed setting, periodic computation of full gradient over all data samples needs to periodically synchronize all clients/devices/machines, which may be impossible or unaffordable. Thus, we expect that ZeroSARAH/D-ZeroSARAH will have a practical impact in distributed and federated learning where full device participation is impractical.

****

VIMPAC: Video Pre-Training via Masked Token Prediction and Contrastive Learning

Hao Tan,Jie Lei,Thomas Wolf,Mohit Bansal

Video understanding relies on perceiving the overall global content and modeling its internal connections (e.g., causality, movement, and spatio-temporal correspondence). To learn these interactions, we apply a mask-then-predict pre-training task on the discretized video tokens generated via VQ-VAE. Unlike language, where the text tokens are more independent, neighboring video tokens typically have strong correlations (e.g., consecutive video frames usually look very similar), and hence uniformly masking individual tokens will make the task too trivial to learn useful representations. To deal with this issue, we propose a block-wise masking strategy where we mask neighboring video tokens in both spatial and temporal domains. We also add an augmentation-free contrastive learning method to further capture the global content by predicting whether the video clips are sampled from the same video. We pre-train our model on uncurated videos and show that our pre-trained model can reach state-of-the-art results on several video understanding datasets (e.g., SSV2, Diving48). Lastly, we provide detailed analyses of the model scalability and pre-training method design.

****

Programmable 3D snapshot microscopy with Fourier convolutional networks

Diptodip Deb,Zhenfei Jiao,Alex Bo-Yuan Chen,Misha Ahrens,Kaspar Podgorski,Srinivas C Turaga

3D snapshot microscopy enables fast volumetric imaging by capturing a 3D volume in a single 2D camera image and performing computational reconstruction. Fast volumetric imaging has a variety of biological applications such as whole brain imaging of rapid neural activity in larval zebrafish. The optimal microscope design for this optical 3D-to-2D encoding is both sample- and task-dependent, with no general solution known. Deep learning based decoders can be combined with a differentiable simulation of an optical encoder for end-to-end optimization of both the deep learning decoder and optical encoder. This technique has been used to engineer local optical encoders for other problems such as depth estimation, 3D particle localization, and lensless photography. However, 3D snapshot microscopy is known to require a highly non-local optical encoder which existing UNet-based decoders are not able to engineer. We show that a neural network architecture based on global kernel Fourier convolutional neural networks can efficiently decode information from multiple depths in a volume, globally encoded across a 3D snapshot image. We show in simulation that our proposed networks succeed in engineering and reconstructing optical encoders for 3D snapshot microscopy where the existing state-of-the-art UNet architecture fails. We also show that our networks outperform the state-of-the-art learned reconstruction algorithms for a computational photography dataset collected on a prototype lensless camera which also uses a highly non-local optical encoding.

****

TimeVAE: A Variational Auto-Encoder for Multivariate Time Series Generation

Abhyuday Desai,Cynthia Freeman,Zuhui Wang,Ian Beaver

Recent work in synthetic data generation in the time-series domain has focused on the use of Generative Adversarial Networks. We propose a novel architecture for synthetically generating time-series data with the use of Variational Auto-Encoders (VAEs). The proposed architecture has several distinct properties: interpretability, ability to encode domain knowledge, and reduced training times. We evaluate data generation quality by similarity and predictability against four multivariate datasets. We experiment with varying sizes of training data to measure the impact of data availability on generation quality for our VAE method as well as several state-of-the-art data generation methods. Our results on similarity tests show that the VAE approach is able to accurately represent the temporal attributes of the original data. On next-step prediction tasks using generated data, the proposed VAE architecture consistently meets or exceeds performance of state-of-the-art data generation methods. While noise reduction may cause the generated data to deviate from original data, we demonstrate the resulting de-noised data can significantly improve performance for next-step prediction using generated data. Finally, the proposed architecture can incorporate domain-specific time-patterns such as polynomial trends and seasonalities to provide interpretable outputs. Such interpretability can be highly advantageous in applications requiring transparency of model outputs or where users desire to inject prior knowledge of time-series patterns into the generative model.

****

Quantifying the Controllability of Coarsely Characterized Networked Dynamical Systems

Nafiseh Ghoroghchian,Rajasekhar Anguluri,Gautam Dasarathy,Stark Draper

We study the controllability of large-scale networked dynamical systems when complete knowledge of network structure is unavailable. In particular, we establish the power of learning community-based representations to understand the ability of a group of control nodes to steer the network to a target state. We are motivated by abundant real-world examples, ranging from power and water systems to brain networks, in which practitioners do not have access to fine-scale knowledge of the network. Rather, knowledge is limited to coarse summaries of network structure. Existing work on "model order reduction" starts with full knowledge of fine-scale structure and derives a coarse-scale (lower-dimensional) model that well-approximates the fine-scale system. In contrast, in this paper the controllability aspects of the coarse system are derived from coarse summaries {\em without} knowledge of the fine-scale structure. We study under what conditions measures of controllability for the (unobserved) fine-scale system can be well approximated by measures of controllability derived from the (observed) coarse-scale system. To accomplish this, we require knowledge of some inherent parametric structure of the fine-scale system that makes this type of inverse problem feasible. To this end, we assume that the underlying fine-scale network is generated by the stochastic block model (SBM) often studied in community detection. We quantify controllability using the ``average controllability'' metric and bound the difference between the controllability of the fine-scale system and that of the coarse-scale system. Our analysis indicates the necessity of underlying structure to make possible the learning of community-based representations, and to be able to quantify accurately the controllability of coarsely characterized networked dynamical systems.

****

HyperTransformer: Attention-Based CNN Model Generation from Few Samples

Andrey Zhmoginov,Max Vladymyrov,Mark Sandler

In this work we propose a HyperTransformer, a transformer based model that generates all weights of a CNN model directly from the support samples. This approach allows to use a high-capacity model for encoding task-dependent variations in the weights of a smaller model. We show for multiple few-shot benchmarks with different architectures and datasets that our method beats or matches that of the traditional learning methods in a few-shot regime. Specifically, we show that for very small target models, our method can generate significantly better performing models than traditional few-shot learning methods. For larger models we discover that applying generation to the last layer only, allows to produce competitive or better results while being end-to-end differentiable. Finally, we extend our approach to semi-supervised regime utilizing unlabeled samples in the support set and further improving few-shot performance in the presence of unlabeled data.

****

Nonparametric Learning of Two-Layer ReLU Residual Units

Zhunxuan Wang,Linyun He,Chunchuan Lyu,Shay B Cohen

We describe an algorithm that learns two-layer residual units using rectified linear unit (ReLU) activation: suppose the input $\mathbf{x}$ is from a distribution with support space $\mathbb{R}^d$ and the ground-truth generative model is a residual unit of this type, given by $\mathbf{y} = \boldsymbol{B}^\ast\left[\left(\boldsymbol{A}^\ast\mathbf{x}\right)^+ + \mathbf{x}\right]$, where ground-truth network parameters $\boldsymbol{A}^\ast \in \mathbb{R}^{d\times d}$ represent a nonnegative full-rank matrix and $\boldsymbol{B}^\ast \in \mathbb{R}^{m\times d}$ is full-rank with $m \geq d$ and for $\boldsymbol{c} \in \mathbb{R}^d$, $[\boldsymbol{c}^{+}]_i = \max\{0, c_i\}$. We design layer-wise objectives as functionals whose analytic minimizers express the exact ground-truth network in terms of its parameters and nonlinearities. Following this objective landscape, learning residual units from finite samples can be formulated using convex optimization of a nonparametric function: for each layer, we first formulate the corresponding empirical risk minimization (ERM) as a positive semi-definite quadratic program (QP), then we show the solution space of the QP can be equivalently determined by a set of linear inequalities, which can then be efficiently solved by linear programming (LP). We further prove the statistical strong consistency of our algorithm, and demonstrate its robustness and sample efficiency through experimental results.

****

Connecting Graph Convolution and Graph PCA

Lingxiao Zhao,Leman Akoglu

Graph convolution operator of the GCN model is originally motivated from a localized first-order approximation of spectral graph convolutions. This work stands on a different view; establishing a mathematical connection between graph convol ution and graph-regularized PCA (GPCA). Based on this connection, the GCN architecture, shaped by stacking graph convolution layers, shares a close relationship with stacking GPCA. We empirically demonstrate that the unsupervised embeddings by GPCA paired with a 1- or 2-layer MLP achieves similar or even better performance than many sophisticated baselines on semi-supervised node classification tasks across five datasets including Open Graph Benchmark. This suggests that the prowess of graph convolution is driven by graph based regularization. In addition, we extend GPCA to the (semi-)supervised setting and show that it is equivalent to GPCA on a graph extended with “ghost” edges between nodes of the same label. Finally, we capitalize on the discovered relationship to design an effective initialization strategy based on stacking GPCA, enabling GCN to converge faster and achieve robust performance at large number of layers.

****

Using Document Similarity Methods to create Parallel Datasets for Code Translation

Mayank Agarwal,Kartik Talamadupula,Fernando Martinez,Stephanie Houde,Michael Muller,John Richards,Steven I Ross,Justin D. Weisz

Translating source code from one programming language to another is a critical, time-consuming task in modernizing legacy applications and codebases. Recent work in this space has drawn inspiration from the software naturalness hypothesis by applying natural language processing techniques towards automating the code translation task. However, due to the paucity of parallel data in this domain, supervised techniques have only been applied to a limited set of popular programming languages. To bypass this limitation, unsupervised neural machine translation techniques have been proposed to learn code translation using only monolingual corpora. In this work, we propose to use document similarity methods to create noisy parallel datasets of code, thus enabling supervised techniques to be applied for automated code translation without having to rely on the availability or expensive curation of parallel code datasets. We explore the noise tolerance of models trained on such automatically-created datasets and show that these models perform comparably to models trained on ground truth for reasonable levels of noise. Finally, we exhibit the practical utility of the proposed method by creating parallel datasets for languages beyond the ones explored in prior work, thus expanding the set of programming languages for automated code translation.

****

Interacting Contour Stochastic Gradient Langevin Dynamics

Wei Deng,Siqi Liang,Botao Hao,Guang Lin,Faming Liang

We propose an interacting contour stochastic gradient Langevin dynamics (ICSGLD) sampler, an embarrassingly parallel multiple-chain contour stochastic gradient Langevin dynamics (CSGLD) sampler with efficient interactions. We show that ICSGLD can be theoretically more efficient than a single-chain CSGLD with an equivalent computational budget. We also present a novel random-field function, which facilitates the estimation of self-adapting parameters in big data and obtains free mode explorations. Empirically, we compare the proposed algorithm with popular benchmark methods for posterior sampling. The numerical results show a great potential of ICSGLD for large-scale uncertainty estimation tasks.

****

NeuPL: Neural Population Learning

Siqi Liu,Luke Marris,Daniel Hennes,Josh Merel,Nicolas Heess,Thore Graepel

Learning in strategy games (e.g. StarCraft, poker) requires the discovery of diverse policies. This is often achieved by iteratively training new policies against existing ones, growing a policy population that is robust to exploit. This iterative approach suffers from two issues in real-world games: a) under finite budget, approximate best-response operators at each iteration needs truncating, resulting in under-trained good-responses populating the population; b) repeated learning of basic skills at each iteration is wasteful and becomes intractable in the presence of increasingly strong opponents. In this work, we propose Neural Population Learning (NeuPL) as a solution to both issues. NeuPL offers convergence guarantees to a population of best-responses under mild assumptions. By representing a population of policies within a single conditional model, NeuPL enable s transfer learning across policies. Empirically, we show the generality, improved performance and efficiency of NeuPL across several test domains. Most interestingly, we show that novel strategies become more accessible, not less, as the neural population expands.

****

PKCAM: Previous Knowledge Channel Attention Module

Eslam Mohamed BAKR,Ahmad A. Al Sallab,Mohsen Rashwan

Attention mechanisms have been explored with CNNs, both across the spatial and channel dimensions.

However, all the existing methods devote the attention modules to capture local interactions from the current feature map only, disregarded the valuable previous knowledge that is acquired by the earlier layers.

This paper tackles the following question: Can one incorporate previous knowledge aggregation while learning channel attention more efficiently? To this end, we propose a Previous Knowledge Channel Attention Module(PKCAM), that captures channel-wise relations across different layers to model the global context.

Our proposed module PKCAM is easily integrated into any feed-forward CNN architectures and trained in an end-to-end fashion with a negligible footprint due to its lightweight property. We validate our novel architecture through extensive experiments on image classification and object detection tasks with different backbones.

Our experiments show consistent improvements in performances against their counterparts. We also conduct experiments that probe the robustness of the learned representations.

****

Reinforcement Learning under a Multi-agent Predictive State Representation Model: Method and Theory

Zhi Zhang,Zhuoran Yang,Han Liu,Pratap Tokekar,Furong Huang

We study reinforcement learning for partially observable multi-agent systems where each agent only has access to its own observation and reward and aims to maximize its cumulative rewards. To handle partial observations, we propose graph-assisted predictive state representations (GAPSR), a scalable multi-agent representation learning framework that leverages the agent connectivity graphs to aggregate local representations computed by each agent. In addition, our representations are readily able to incorporate dynamic interaction graphs and kernel space embeddings of the predictive states, and thus have strong flexibility and representation power.

Based on GAPSR, we propose an end-to-end MARL algorithm that simultaneously infers the predictive representations and uses the representations as the input of a policy optimization algorithm. Empirically, we demonstrate the efficacy of the proposed algorithm provided on both a MAMuJoCo robotic learning experiment and a multi-agent particle learning environment.

****

Meta-Referential Games to Learn Compositional Learning Behaviours

Kevin Yandoka Denamganai,Sondess Missaoui,James Alfred Walker

Referring to compositional learning behaviours as the ability to learn to generalise compositionally from a limited set of stimuli, that are combinations of supportive stimulus components, to a larger set of novel stimuli, i.e. novel combinations of those same stimulus components, we acknowledge compositional learning behaviours as a valuable feat of intelligence that human beings often rely on, and assume their collaborative partners to use similarly. In order to build artificial agents able to collaborate with human beings, we propose a novel benchmark to investigate state-of-the-art artificial agents abilities to exhibit compositional learning behaviours. We provide baseline results on the single-agent tasks of learning compositional learning behaviours, using state-of-the-art RL agents, and show that our proposed benchmark is a compelling challenge that we hope will spur the research community towards developing more capable artificial agents.

****

A New Perspective on "How Graph Neural Networks Go Beyond Weisfeiler-Lehman?"

Asiri Wijesinghe,Qing Wang

We propose a new perspective on designing powerful Graph Neural Networks (GNNs). In a nutshell, this enables a general solution to inject structural properties of graphs into a message-passing aggregation scheme of GNNs. As a theoretical basis, we develop a new hierarchy of local isomorphism on neighborhood subgraphs. Then, we theoretically characterize how message-passing GNNs can be designed to be more expressive than the Weisfeiler Lehman test. To elaborate this characterization, we propose a novel neural model, called GraphSNN, and prove that this model is strictly more expressive than the Weisfeiler Lehman test in distinguishing graph structures. We empirically verify the strength of our model on different graph learning tasks. It is shown that our model consistently improves the state-of-the-art methods on the benchmark tasks without sacrificing computational simplicity and efficiency.

****

DeSKO: Stability-Assured Robust Control with a Deep Stochastic Koopman Operator

Minghao Han,Jacob Euler-Rolle,Robert K. Katzschmann

The Koopman operator theory linearly describes nonlinear dynamical systems in a high-dimensional functional space and it allows to apply linear control methods to highly nonlinear systems. However, the Koopman operator does not account for any uncertainty in dynamical systems, causing it to perform poorly in real-world applications.

Therefore, we propose a deep stochastic Koopman operator (DeSKO) model in a robust learning control framework to guarantee stability of nonlinear stochastic systems. The DeSKO model captures a dynamical system's uncertainty by inferring a distribution of observables. We use the inferred distribution to design a robust, stabilizing closed-loop controller for a dynamical system. Modeling and control experiments on several advanced control benchmarks show that our framework is more robust and scalable than state-of-the-art deep Koopman operators and reinforcement learning methods. Tested control benchmarks include a soft robotic arm, a legged robot, and a biological gene regulatory network. We also demonstrate that this robust control method resists previously unseen uncertainties, such as external disturbances, with a magnitude of up to five times the maximum control input. Our approach opens up new possibilities in learning control for high-dimensional nonlinear systems while robustly managing internal or external uncertainty.

****

Where can quantum kernel methods make a big difference?

Muhao Guo,Yang Weng

The classification problem is a core problem of supervised learning, which is widely present in our life. As a class of algorithms for pattern analysis, Kernel methods have been widely and effectively applied to classification problems. However, when very complex patterns are encountered, the existing kernel methods are powerless. Recent studies have shown that quantum kernel methods can effectively handle some classification problems of complex patterns that classical kernel methods cannot handle. However, this does not mean that quantum kernel methods are better than classical kernel methods in all cases. It is still unclear under what circumstances quantum kernel methods can realize their great potential. In this paper, by exploring and summarizing the essential differences between quantum kernel functions and classical kernel functions, we propose a criterion based on inter-class and intra-class distance and geometric properties to determine under what circumstances quantum kernel methods will be superior. We validate our method with toy examples and multiple real datasets from Qiskit and Kaggle. The experiments show that our method can be used as a valid determination method.

****

Learning to Efficiently Sample from Diffusion Probabilistic Models

Daniel Watson,Jonathan Ho,Mohammad Norouzi,William Chan

Denoising Diffusion Probabilistic Models (DDPMs) have emerged as a powerful family of generative models that, yielding high-fidelity samples and competitive log-likelihoods across a range of domains, including image and speech synthesis. Key advantages of DDPMs include ease of training, in contrast to generative advers arial networks, and speed of generation, in contrast to autoregressive models. However, DDPMs typically require hundreds-to-thousands of steps to generate a high fidelity sample, making them prohibitively expensive for high dimensional problems. Fortunately, DDPMs allow trading generation speed for sample quality through adjusting the number of refinement steps during inference. Prior work has been successful in improving generation speed through handcrafting the time schedule through trial and error. In our work, we view the selection of the inference time schedules as an optimization problem, and introduce an exact dynamic programming algorithm that finds the log-likelihood-optimal discrete time schedules for any pre-trained DDPM. Our method exploits the fact that the evidence lower bound (ELBO) can be decomposed into separate KL divergence terms, and given any computation budget, we discover the time schedule that maximizes the training ELBO exactly. Our method is efficient, has no hyper-parameters of its own, and can be applied to any pre-trained DDPM with no retraining. We discover inference time schedules requiring as few as 32 refinement steps, while sacrificing less than 0.1 bits per dimension compared to the default 4,000 steps used on an ImageNet 64x64 model.

****

An Investigation into the Role of Author Demographics in ICLR Participation and Review

Keshav Ganapathy,Emily Liu,Zain Zarger,Gowthami Somepalli,Micah Goldblum,Tom Goldstein

As machine learning conferences grow rapidly, many are concerned that individuals will be left behind on the basis of traits such as gender and geography. We leverage historic ICLR submissions from 2017 to 2021 to investigate the impact of gender and country of origin both on representation and paper review outcomes at ICLR. We also study various hypotheses that could explain gender representation disparities at ICLR, with a focus on factors that impact the likelihood of an author returning to the conference in consecutive years. Finally, we probe the effects of paper topic on the review process and perform a study on how the inclusion of theorems and the number of co-authors impact the success of papers in the review process.

****

Effects of Data Geometry in Early Deep Learning

Saket Tiwari,George Konidaris

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 number of linear regions and the distance to boundaries of these linear regions 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 how our results hold upon changing neural network architectures. We further demonstrate how the complexity of linear regions changes on the low dimensional manifold of images as training progresses, using the MetFaces dataset.

****

Design in the Dark: Learning Deep Generative Models for De Novo Protein Design

Lewis Moffat,Shaun M. Kandathil,David T. Jones

The design of novel protein sequences is providing paths towards the development of novel therapeutics and materials.

Generative modelling approaches to design are emerging and to date have required conditioning on 3D protein structure-derived information, and unconditional models of protein sequences have so far performed poorly.

Thus, it is unknown if unconditional generative models can learn a distribution of sequences that captures structure information without it being explicitly provided, and so be of use in important tasks like de novo protein sequence design, where it is not possible to condition on structure.

Here, we demonstrate that it is possible to use unconditioned generative models to produce realistic samples of protein sequences.

We progressively grow a dataset of over half a million synthetic sequences for training autoregressive language models, using an iterative framework we call DARK.

It begins by training an autoregressive model on an initial sample of synthetic sequences, sampling from it, and refining the samples thus generated, which are then used for subsequent rounds of training.

Using the confidence measures provided by AlphaFold and other measures of sample quality, we show that our approach matches or exceeds the performance of prior methods that use weak conditioning on explicit structural information, and improves after each iteration of DARK.

Crucially, the DARK framework and the trained models are entirely unsupervised; strong structural signal is an objective, but no model is ever conditioned on any specific structural state.

The trained model indirectly learns to incorporate a structural signal into its learned sequence distribution, as this signal is strongly represented in the makeup of the training set at each step.

Our work demonstrates a way of unconditionally sampling sequences and structures jointly, and in an unsupervised way.

****

Information-theoretic stochastic contrastive conditional GAN: InfoSCC-GAN

Vitaliy Kinakh,Mariia Drozdova,Guillaume Quétant,Svyatoslav Voloshynovskyy,Tobias GOLLING

Conditional generation is a subclass of generative problems when the output of generation is conditioned by a class attributes’ information. In this paper, we present a new stochastic contrastive conditional generative adversarial network (InfoSCC-GAN) with explorable latent space. The InfoSCC-GAN architecture is based on an unsupervised contrastive encoder built on the InfoNCE paradigm, attributes' classifier, and stochastic EigenGAN generator.

We propose two approaches for selecting the class attributes: external attributes from the dataset annotations and internal attributes from the clustered latent space of the encoder. We propose a novel training method based on a generator regularization using external or internal attributes every n-th iteration using the pre-trained contrastive encoder and pre-trained attributes’ classifier. The proposed InfoSCC-GAN is derived from an information-theoretic formulation of mutual information maximization between the input data and latent space representation for the encoder and the latent space and generated data for the decoder. Thus, we demonstrate a link between the training objective functions and the above information-theoretic formulation. The experimental results show that InfoSCC-GAN outperforms vanilla EigenGAN in image generation on several popular datasets, yet providing an interpretable latent space. In addition, we investigate the impact of regularization techniques and each part of the system by performing an ablation study. Finally, we demonstrate that thanks to the stochastic EigenGAN generator, the proposed framework enjoys a truly stochastic generation in contrast to vanilla deterministic GANs yet with the independent training of an encoder, a classifier, and a generator.

The code, supplementary materials, and demos are available \url{https://anonymous.4open.science/r/InfoSCC-GAN-D113}

****

Defending Backdoor Data Poisoning Attacks by Using Noisy Label Defense Algorithm

Boyang Liu,Zhuangdi Zhu,Pang-Ning Tan,Jiayu Zhou

Training deep neural networks with data corruption is a challenging problem. One example of such corruption is the backdoor data poisoning attack, in which an adversary strategically injects a backdoor trigger to a small fraction of the training data to subtly compromise the training process. Consequently, the trained deep neural network would misclassify testing examples that have been corrupted by the same trigger. While the label of the data could be changed to arbitrary values by an adversary, the extent of corruption injected to the feature values a re strictly limited in order to keep the backdoor attack in disguise, which leads to a resemblance between the backdoor attack and a milder attack that involves only noisy labels. In this paper, we investigate an intriguing question: Can we leverage algorithms that defend against noisy labels corruptions to defend against general backdoor attacks? We first discuss the limitations of directly using the noisy-label defense algorithms to defend against backdoor attacks. Next, we propose a meta-algorithm that transforms an existing noisy label defense algorithm to one that protects against backdoor attacks. Extensive experiments on different types of backdoor attacks show that, by introducing a lightweight alteration for minimax optimization to the existing noisy-label defense algorithms, the robustness against backdoor attacks can be substantially improved, while the intial form of those algorithms would fail in presence of a backdoor attacks.

****

Degradation Attacks on Certifiably Robust Neural Networks

Klas Leino,Chi Zhang,Ravi Mangal,Matt Fredrikson,Bryan Parno,Corina Pasareanu

Certifiably robust neural networks employ provable run-time defenses against adversarial examples by checking if the model is locally robust at the input under evaluation. We show through examples and experiments that these defenses are inherently over-cautious. Specifically, they flag inputs for which local robustness checks fail, but yet that are not adversarial; i.e., they are classified consistently with all valid inputs within a distance of ϵ. As a result, while a norm-bounded adversary cannot change the classification of an input, it can use norm-bounded changes to degrade the utility of certifiably robust networks by forcing them to reject otherwise correctly classifiable inputs. We empirically demonstrate the efficacy of such attacks against state-of-the-art certifiable defenses.

****

Neural Network Approximation based on Hausdorff distance of Tropical Zonotopes

Panagiotis Misiakos,Georgios Smyrnis,George Retsinas,Petros Maragos

In this work we theoretically contribute to neural network approximation by providing a novel tropical geometrical viewpoint to structured neural network compression. In particular, we show that the approximation error between two neural networks with ReLU activations and one hidden layer depends on the Hausdorff distance of the tropical zonotopes of the networks. This theorem comes as a first step towards a purely geometrical interpretation of neural network approximation. Based on this theoretical contribution, we propose geometrical methods that employ the K-means algorithm to compress the fully connected parts of ReLU activated deep neural networks. We analyze the error bounds of our algorithms theoretically based on our approximation theorem and evaluate them empirically on neural network compression. Our experiments follow a proof-of-concept strategy and indicate that our geometrical tools achieve improved performance over relevant tropical geometry techniques and can be competitive against non-tropical methods.

****

Mixture Representation Learning with Coupled Autoencoders

Yeganeh Marghi,Rohan Gala,Uygar Sümbül

Latent representations help unravel complex phenomena. While continuous latent variables can be efficiently inferred, fitting mixed discrete-continuous models remains challenging despite recent progress, especially when the discrete factor dimensionality is large. A pressing application for such mixture representations is the analysis of single-cell omic datasets to understand neuronal diversity and its molecular underpinnings. Here, we propose an unsupervised variational framework using multiple interacting networks called cpl-mixVAE that significantly outperforms state-of-the-art in high-dimensional discrete settings. cpl-mixVAE introduces a consensus constraint on discrete factors of variability across the networks, which regularizes the mixture representations at the time of training. We justify the use of this framework with theoretical results and validate it with experiments on benchmark datasets. We demonstrate that our approach discovers interpretable discrete and continuous variables describing neuronal identity in two single-cell RNA sequencing datasets, each profiling over a hundred cortical neuron types.

****

Generalization to Out-of-Distribution transformations

Shanka Subhra Mondal,Zack Dulberg,Jonathan Cohen

Humans understand a set of canonical geometric transformations (such as translation, rotation and scaling) that support generalization by being untethered to any specific object. We explored inductive biases that allowed artificial neural networks to learn these transformations in pixel space in a way that could generalize out-of-distribution (OOD). Unsurprisingly, we found that convolution and high training diversity were important contributing factors to OOD generalization of translation to untrained shapes, sizes, time-points and locations, however these weren’t sufficient for rotation and scaling. To remedy this we show that two more principle components are needed 1) iterative training where outputs are fed back as inputs 2) applying convolutions after conversion to log-polar space. We propose POLARAE which exploits all four components and outperforms standard convolutional autoencoders and variational autoencoders trained iteratively with high diversity wrt OOD generalization to larger shapes in larger grids and new locations.

****

Learning Towards The Largest Margins

Xiong Zhou,Xianming Liu,Deming Zhai,Junjun Jiang,Xin Gao,Xiangyang Ji

One of the main challenges for feature representation in deep learning-based classification is the design of appropriate loss functions that exhibit strong discriminative power. The classical softmax loss does not explicitly encourage discriminative learning of features. A popular direction of research is to incorporate margins in well-established losses in order to enforce extra intra-class compactness and inter-class separability, which, however, were developed through heuristic means, as opposed to rigorous mathematical principles. In this work, we attempt to address this limitation by formulating the principled optimization objective as learning towards the largest margins. Specifically, we firstly propose to employ the class margin as the measure of inter-class separability, and the sample margin as the measure of intra-class compactness. Accordingly, to encourage discriminative representation of features, the loss function should promote the largest possible margins for both classes and samples. Furthermore, we derive a generalized margin softmax loss to draw general conclusions for the existing margin-based losses. Not only does this principled framework offer new perspectives to understand and interpret existing margin-based losses, but it also provides new insights that can guide the design of new tools, including \textit{sample margin regularization} and \textit{largest margin softmax loss} for class balanced cases, and \textit{zero centroid regularization} for class imbalanced cases. Experimental results demonstrate the effectiveness of our strategy for multiple tasks including visual classification, imbalanced classification, person re-identification, and face verification.

****

Adversarial Support Alignment

Shangyuan Tong,Timur Garipov,Yang Zhang,Shiyu Chang,Tommi S. Jaakkola

We study the problem of aligning the supports of distributions. Compared to the existing work on distribution alignment, support alignment does not require the densities to be matched. We propose symmetric support difference as a divergence measure to quantify the mismatch between supports. We show that select discriminators (e.g. discriminator trained for Jensen-Shannon divergence) are able to map support differences as support differences in their one-dimensional output space. Following this result, our method aligns supports by minimizing a symmetrized relaxed optimal transport cost in the discriminator 1D space via an adversarial process. Furthermore, we show that our approach can be viewed as a limit of existing notions of alignment by increasing transportation assignment tolerance. We quantitatively evaluate the method across domain adaptation tasks with shifts in label distributions. Our experiments show that the proposed method is more ro bust against these shifts than other alignment-based baselines.

****

TorchGeo: deep learning with geospatial data

Adam J Stewart,Caleb Robinson,Isaac A Corley,Anthony Ortiz,Juan M Lavista Ferres,Arindam Banerjee

Remotely sensed geospatial data are critical for earth observation applications including precision agriculture, urban planning, disaster monitoring and response, and climate change research, among others. Deep learning methods are particularly promising for modeling many earth observation tasks given the success of deep neural networks in similar computer vision tasks and the sheer volume of remotely sensed imagery available. However, the variance in data collection methods and handling of geospatial metadata make the application of deep learning methodology to remotely sensed data nontrivial. For example, satellite imagery often includes additional spectral bands beyond red, green, and blue and must be joined to other geospatial data sources that can have differing coordinate systems, bounds, and resolutions. To help realize the potential of deep learning for remote sensing applications, we introduce TorchGeo, a Python library for integrating geospatial data into the PyTorch deep learning ecosystem. TorchGeo provides data loaders for a variety of benchmark datasets, composable datasets for generic geospatial data sources, samplers for geospatial data, and transforms that work with multispectral imagery. TorchGeo is also the first library to provide pre-trained models for multispectral satellite imagery, allowing for advances in transfer learning on downstream earth observation tasks with limited labeled data. We use TorchGeo to create reproducible benchmark results on existing datasets, benchmark our proposed method for preprocessing geospatial imagery on-the-fly, and investigate the differences between ImageNet pre-training and in-domain self-supervised pre-training on model performance across several datasets. We aim for TorchGeo to become a new standard for reproducibility and for driving progress at the intersection of deep learning and remotely sensed geospatial data.

****

Patch-Fool: Are Vision Transformers Always Robust Against Adversarial Perturbations?

Yonggan Fu,Shunyao Zhang,Shang Wu,Cheng Wan,Yingyan Lin

Vision transformers (ViTs) have recently set off a new wave in neural architecture design thanks to their record-breaking performance in various vision tasks. In parallel, to fulfill the goal of deploying ViTs into real-world vision applications, their robustness against potential malicious attacks has gained increasing attention. In particular, recent works show that ViTs are more robust against adversarial attacks as compared with convolutional neural networks (CNNs), and conjecture that this is because ViTs focus more on capturing global interactions among different input/feature patches, leading to their improved robustness to local perturbations imposed by adversarial attacks. In this work, we ask an intriguing question: "Under what kinds of perturbations do ViTs become more vulnerable learners compared to CNNs?" Driven by this question, we first conduct a comprehensive experiment regarding the robustness of both ViTs and CNNs under various existing adversarial attacks to understand the underlying reason favoring their robustness. Based on the drawn insights, we then propose a dedicated attack framework, dubbed Patch-Fool, that fools the self-attention mechanism by attacking its basic component (i.e., a single patch) with a series of attention-aware optimization techniques. Interestingly, our Patch-Fool framework shows for the first time that ViTs are not necessarily more robust than CNNs against adversarial perturbations. In particular, we find that ViTs are more vulnerable learners compared with CNNs against our Patch-Fool attack which is consistent across extensive experiments, and the observations from Sparse/Mild Patch-Fool, two variants of Patch-Fool, indicate an intriguing insight that the perturbation density and strength on each patch seem to be the key factors that influence the robustness ranking between ViTs and CNNs. It can be expected that our Patch-Fool framework will shed light on both future architecture designs and training schemes for robustifying ViTs towards their real-world deployment. Our codes are available at https://github.com/RICE-EIC/Patch-Fool.

****

A Survey on Evidential Deep Learning For Single-Pass Uncertainty Estimation
Dennis Thomas Ulmer

Popular approaches for quantifying predictive uncertainty in deep neural networks often involve a set of weights or models, for instance via ensembling or Monte Carlo Dropout. These techniques usually produce overhead by having to train multiple model instances or do not produce very diverse predictions. This survey aims to familiarize the reader with an alternative class of models based on the concept of Evidential Deep Learning: For unfamiliar data, they admit “what they don’t know” and fall back onto a prior belief. Furthermore, they allow uncertainty estimation in a single model and forward pass by parameterizing distributions over distributions. This survey recapitulates existing works, focusing on the implementation in a classification setting. Finally, we survey the application of the same paradigm to regression problems. We also provide a reflection on the strengths and weaknesses of the mentioned approaches compared to existing ones and provide the most central theoretical results in order to inform future research.

****

AdaMatch: A Unified Approach to Semi-Supervised Learning and Domain Adaptation
David Berthelot,Rebecca Roelofs,Kihyuk Sohn,Nicholas Carlini,Alexey Kurakin

We extend semi-supervised learning to the problem of domain adaptation to learn significantly higher-accuracy models that train on one data distribution and test on a different one. With the goal of generality, we introduce AdaMatch, a unified solution for unsupervised domain adaptation (UDA), semi-supervised learning (SSL), and semi-supervised domain adaptation (SSDA). In an extensive experimental study, we compare its behavior with respective state-of-the-art techniques from SSL, SSDA, and UDA and find that AdaMatch either matches or significantly exceeds the state-of-the-art in each case using the same hyper-parameters regardless of the dataset or task. For example, AdaMatch nearly doubles the accuracy compared to that of the prior state-of-the-art on the UDA task for DomainNet and even exceeds the accuracy of the prior state-of-the-art obtained with pre-training by 6.4% when AdaMatch is trained completely from scratch. Furthermore, by providing AdaMatch with just one labeled example per class from the target domain (i.e., the SSDA setting), we increase the target accuracy by an additional 6.1%, and with 5 labeled examples, by 13.6%.

****

Complete Verification via Multi-Neuron Relaxation Guided Branch-and-Bound
Claudio Ferrari,Mark Niklas Mueller,Nikola Jovanovi■,Martin Vechev

State-of-the-art neural network verifiers are fundamentally based on one of two paradigms: either encoding the whole verification problem via tight multi-neuron convex relaxations or applying a Branch-and-Bound (BaB) procedure leveraging imprecise but fast bounding methods on a large number of easier subproblems. The former can capture complex multi-neuron dependencies but sacrifices completeness due to the inherent limitations of convex relaxations. The latter enables complete verification but becomes increasingly ineffective on larger and more challenging networks. In this work, we present a novel complete verifier which combines the strengths of both paradigms: it leverages multi-neuron relaxations to drastically reduce the number of subproblems generated during the BaB process and an efficient GPU-based dual optimizer to solve the remaining ones. An extensive evaluation demonstrates that our verifier achieves a new state-of-the-art on both established benchmarks as well as networks with significantly higher accuracy than previously considered. The latter result (up to 28% certification gains) indicates meaningful progress towards creating verifiers that can handle practically relevant networks.

****

DATA-DRIVEN EVALUATION OF TRAINING ACTION SPACE FOR REINFORCEMENT LEARNING
Rajat Ghosh,Debojyoti Dutta,Aroosh Sohi,Akshay Khole

Training action space selection for reinforcement learning (RL) is conflict-prone due to complex state-action relationships. To address this challenge, this paper proposes a Shapely-inspired methodology for training action space categorization and ranking. To reduce exponential-time Shapely computations, the methodolog y includes a Monte Carlo simulation to avoid unnecessary explorations. The effectiveness of the methodology is illustrated using a cloud infrastructure resource tuning case study. It reduces the search space by 80% and categorizes the training action sets into dispensable and indispensable groups. Additionally, it ranks different training actions to facilitate superior RL model performance and lower cost. The proposed data-driven methodology is extensible to different domains, use cases, and machine learning algorithms.

****

Kokoyi: Executable LaTeX for End-to-end Deep Learning

Minjie Wang,Haoming Lu,Yu Gai,Lesheng Jin,Zihao Ye,Zheng Zhang

Despite substantial efforts from the deep learning system community to relieve researchers and practitioners from the burden of implementing models with ever-growing complexity, a considerable lingual gap remains between developing models in the language of mathematics and implementing them in the languages of computer. The mission of Kokoyi is to close this gap by enabling automatic translation of mathematics into efficient implementations, thereby making math-in-codes and math-in-model consistent. This paper presents our first step towards the goal: kokoyi-lang, a programming language with the syntax of LaTeX and the semantics of deep learning mathematics, and a prototype kokoyi-lang compiler and runtime supporting advanced optimizations such as auto-batching. Kokoyi is integrated with Jupyter Notebook, and will be released in open-source.

****

Practical Adversarial Attacks on Brain--Computer Interfaces

Rodolfo Octavio Siller Quintanilla,Xiaying Wang,Michael Hersche,Luca Benini,Gagandeep Singh

Deep learning has been widely employed in brain--computer interfaces (BCIs) to decode a subject's intentions based on recorded brain activities enabling direct interaction with computers and machines. BCI systems play a crucial role in motor rehabilitation and have recently experienced a significant market boost as consumer-grade products. Recent studies have shown that deep learning-based BCIs are vulnerable to adversarial attacks. Failures in such systems might cause medical misdiagnoses, physical harm, and financial damages, hence it is of utmost importance to analyze and understand in-depth, potential malicious attacks to develop countermeasures. In this work, we present the first study that analyzes and models adversarial attacks based on physical domain constraints in EEG-based BCIs. Specifically, we assess the robustness of EEGNet which is the current state-of-the-art network for embedded BCIs. We propose new methods to induce denial-of-service attacks and incorporate domain-specific insights and constraints to accomplish two key goals: (i) create smooth adversarial attacks that are physiologically plausible; (ii) consider the realistic case where the attack happens at the origin of the signal acquisition and it propagates on the human head. Our results show that EEGNet is significantly vulnerable to adversarial attacks with an attack success rate of more than 50\%. With our work, we want to raise awareness and incentivize future developments of proper countermeasures.

****

Input Dependent Sparse Gaussian Processes

Bahram Jafrasteh,Carlos Villacampa-Calvo,Daniel Hernández-Lobato

Gaussian Processes (GPs) are Bayesian models that provide uncertainty estimates associated to the predictions made. They are also very flexible due to their non-parametric nature. Nevertheless, GPs suffer from poor scalability as the number of training instances N increases. More precisely, they have a cubic cost with respect to N. To overcome this problem, sparse GP approximations are often used, where a set of $M \ll N$ inducing points is introduced during training. The location of the inducing points is learned by considering them as parameters of an approximate posterior distribution q. Sparse GPs, combined with variational inference for inferring q, reduce the training cost of GPs to $\mathcal{O}(M^3)$. Critically, the inducing points determine the flexibility of the model and they are often located in regions of the input space where the latent function changes. A limitation is, however, that for some learning tasks a large number of inducing points may be required to obtain a good prediction performance. To ad dress this limitation, we propose here to amortize the computation of the inducing points locations, as well as the parameters of the variational posterior approximation q. For this, we use a neural network that receives the observed data as an input and outputs the inducing points locations and the parameters of q. We evaluate our method in several experiments, showing that it performs similar or better than other state-of-the-art sparse variational GP approaches. However, with our method the number of inducing points is reduced drastically due to their dependency on the input data. This makes our method scale to larger datasets and have faster training and prediction times.

****

GreaseLM: Graph REASoning Enhanced Language Models

Xikun Zhang,Antoine Bosselut,Michihiro Yasunaga,Hongyu Ren,Percy Liang,Christopher D Manning,Jure Leskovec

Answering complex questions about textual narratives requires reasoning over both stated context and the world knowledge that underlies it. However, pretrained language models (LM), the foundation of most modern QA systems, do not robustly represent latent relationships between concepts, which is necessary for reasoning. While knowledge graphs (KG) are often used to augment LMs with structured representations of world knowledge, it remains an open question how to effectively fuse and reason over the KG representations and the language context, which provides situational constraints and nuances. In this work, we propose GreaseLM, a new model that fuses encoded representations from pretrained LMs and graph neural networks over multiple layers of modality interaction operations. Information from both modalities propagates to the other, allowing language context representations to be grounded by structured world knowledge, and allowing linguistic nuances (e.g., negation, hedging) in the context to inform the graph representations of knowledge. Our results on three benchmarks in the commonsense reasoning (i.e., CommonsenseQA, OpenbookQA) and medical question answering (i.e., MedQA-USMLE) domains demonstrate that GreaseLM can more reliably answer questions that require reasoning over both situational constraints and structured knowledge, even outperforming models 8x larger.

****

Learning Fast Samplers for Diffusion Models by Differentiating Through Sample Quality

Daniel Watson,William Chan,Jonathan Ho,Mohammad Norouzi

Diffusion models have emerged as an expressive family of generative models rivaling GANs in sample quality and autoregressive models in likelihood scores. Standard diffusion models typically require hundreds of forward passes through the model to generate a single high-fidelity sample. We introduce Differentiable Diffusion Sampler Search (DDSS): a method that optimizes fast samplers for any pre-trained diffusion model by differentiating through sample quality scores. We also present Generalized Gaussian Diffusion Models (GGDM), a family of flexible non-Markovian samplers for diffusion models. We show that optimizing the degrees of freedom of GGDM samplers by maximizing sample quality scores via gradient descent leads to improved sample quality. Our optimization procedure backpropagates through the sampling process using the reparametrization trick and gradient rematerialization. DDSS achieves strong results on unconditional image generation across various datasets (e.g., FID scores on LSUN church 128x128 of 11.6 with only 10 inference steps, and 4.82 with 20 steps, compared to 51.1 and 14.9 with strongest DDPM/DDIM baselines). Our method is compatible with any pre-trained diffusion model without fine-tuning or re-training required.

****

Distribution Compression in Near-Linear Time

Abhishek Shetty,Raaz Dwivedi,Lester Mackey

In distribution compression, one aims to accurately summarize a probability distribution $\mathbb{P}$ using a small number of representative points. Near-optimal thinning procedures achieve this goal by sampling n points from a Markov chain and identifying $\sqrt{n}$ points with $\widetilde{\mathcal{O}}(1/\sqrt{n})$ discrepancy to $\mathbb{P}$. Unfortunately, these algorithms suffer from quadratic or super-quadratic runtime in the sample size n. To address this deficiency

, we introduce Compress++, a simple meta-procedure for speeding up any thinning algorithm while suffering at most a factor of 4 in error. When combined with the quadratic-time kernel halving and kernel thinning algorithms of Dwivedi and Mackey (2021), Compress++ delivers $\sqrt{n}$ points with $\mathcal{O}(\sqrt{\log n/n})$ integration error and better-than-Monte-Carlo maximum mean discrepancy in $\mathcal{O}(n \log^3 n)$ time and $\mathcal{O}(\sqrt{n} \log^2 n)$ space. Moreover, Compress++ enjoys the same near-linear runtime given any quadratic-time input and reduces the runtime of super-quadratic algorithms by a square-root factor. In our benchmarks with high-dimensional Monte Carlo samples and Markov chains targeting challenging differential equation posteriors, Compress++ matches or nearly matches the accuracy of its input algorithm in orders of magnitude less time.

****

Scalable Robust Federated Learning with Provable Security Guarantees

Andrew Liu,Jacky Y. Zhang,Nishant Kumar,Dakshita Khurana,Oluwasanmi O Koyejo

Federated averaging, the most popular aggregation approach in federated learning, is known to be vulnerable to failures and adversarial updates from clients that wish to disrupt training. While median aggregation remains one of the most popular alternatives to improve training robustness, the naive combination of median and secure multi-party computation (MPC) is unscalable. To this end, we propose an efficient approximate median aggregation with MPC privacy guarantees on the multi-silo setting, e.g., across hospitals, with two semi-honest non-colluding servers. The proposed method protects the confidentiality of client gradient updates against both semi-honest clients and servers. Asymptotically, the cost of our approach scales only linearly with the number of clients, whereas the naive MPC median scales quadratically. Moreover, we prove that the convergence of the proposed federated learning method is robust to a wide range of failures and attacks. Empirically, we show that our method inherits the robustness properties of the median while converging faster than the naive MPC median for even a small number of clients.

****

Capturing Structural Locality in Non-parametric Language Models

Frank F. Xu,Junxian He,Graham Neubig,Vincent Josua Hellendoorn

Structural locality is a ubiquitous feature of real-world datasets, wherein data points are organized into local hierarchies. Some examples include topical clusters in text or project hierarchies in source code repositories. In this paper, we explore utilizing this structural locality within non-parametric language models, which generate sequences that reference retrieved examples from an external source. We propose a simple yet effective approach for adding locality information into such models by adding learned parameters that improve the likelihood of retrieving examples from local neighborhoods. Experiments on two different domains, Java source code and Wikipedia text, demonstrate that locality features improve model efficacy over models without access to these features, with interesting differences. We also perform an analysis of how and where locality features contribute to improving performance and why the traditionally used contextual similarity metrics alone are not enough to grasp the locality structure.

****

Audio Lottery: Speech Recognition Made Ultra-Lightweight, Noise-Robust, and Transferable

Shaojin Ding,Tianlong Chen,Zhangyang Wang

Lightweight speech recognition models have seen explosive demands owing to a growing amount of speech-interactive features on mobile devices. Since designing such systems from scratch is non-trivial, practitioners typically choose to compress large (pre-trained) speech models. Recently, lottery ticket hypothesis reveals the existence of highly sparse subnetworks that can be trained in isolation without sacrificing the performance of the full models. In this paper, we investigate the tantalizing possibility of using lottery ticket hypothesis to discover lightweight speech recognition models, that are (1) robust to various noise existing in speech; (2) transferable to fit the open-world personalization; and 3) co mpatible with structured sparsity. We conducted extensive experiments on CNN-LSTM, RNN-Transducer, and Transformer models, and verified the existence of highly sparse winning tickets that can match the full model performance across those backbones. We obtained winning tickets that have less than 20% of full model weights on all backbones, while the most lightweight one only keeps 4.4% weights. Those winning tickets generalize to structured sparsity with no performance loss, and transfer exceptionally from large source datasets to various target datasets. Perhaps most surprisingly, when the training utterances have high background noises, the winning tickets even substantially outperform the full models, showing the extra bonus of noise robustness by inducing sparsity. Codes are available at https://github.com/VITA-Group/Audio-Lottery.

****

Neural Bootstrapping Attention for Neural Processes

Minsub Lee,Junhyun Park,Sojin Jang,Chanhui Lee,Hyungjoo Cho,Minsuk Shin,Sungbin Lim

Neural Processes (NP) learn to fit a broad class of stochastic processes with neural networks. Modeling functional uncertainty is an important aspect of learning stochastic processes. Recently, Bootstrapping (Attentive) Neural Processes (B(A)NP) propose a bootstrap method to capture the functional uncertainty which can replace the latent variable in (Attentive) Neural Processes ((A)NP), thus overcoming the limitations of Gaussian assumption on the latent variable. However, B(A)NP conduct bootstrapping in a non-parallelizable and memory-inefficient way and fail to capture diverse patterns in the stochastic processes. Furthermore, we found that ANP and BANP both tend to overfit in some cases. To resolve these problems, we propose an efficient and easy-to-implement approach, Neural Bootstrapping Attentive Neural Processes (NeuBANP). NeuBANP learns to generate the bootstrap distribution of random functions by injecting multiple random weights into the encoder and the loss function. We evaluate our models in benchmark experiments including Bayesian optimization and contextual multi-armed bandit. NeuBANP achieves state-of-the-art performance in both of the sequential decision-making tasks, and this empirically shows that our method greatly improves the quality of functional uncertainty modeling.

****

Learning meta-features for AutoML

Herilalaina Rakotoarison,Louisot Milijaona,Andry RASOANAIVO,Michele Sebag,Marc Schoenauer

This paper tackles the AutoML problem, aimed to automatically select an ML algorithm and its hyper-parameter configuration most appropriate to the dataset at hand. The proposed approach, MetaBu, learns new meta-features via an Optimal Transport procedure, aligning the manually designed \mf s with the space of distributions on the hyper-parameter configurations. MetaBu meta-features, learned once and for all, induce a topology on the set of datasets that is exploited to define a distribution of promising hyper-parameter configurations amenable to AutoML. Experiments on the OpenML CC-18 benchmark demonstrate that using MetaBu meta-features boosts the performance of state of the art AutoML systems, AutoSklearn (Feurer et al. 2015) and Probabilistic Matrix Factorization (Fusi et al. 2018). Furthermore, the inspection of MetaBu meta-features gives some hints into when an ML algorithm does well. Finally, the topology based on MetaBu meta-features enables to estimate the intrinsic dimensionality of the OpenML benchmark w.r.t. a given ML algorithm or pipeline. The source code is available at https://github.com/luxusg1/metabu.

****

Adversarially Robust Models may not Transfer Better: Sufficient Conditions for Domain Transferability from the View of Regularization

Xiaojun Xu,Jacky Y. Zhang,Evelyn Ma,Danny Son,Oluwasanmi O Koyejo,Bo Li

Machine learning (ML) robustness and generalization are fundamentally correlated: they essentially concern about data distribution shift under adversarial and natural settings, respectively. Thus, it is critical to uncover their underlying connections to tackle one based on the other. On the one hand, recent studies show that more robust (adversarially trained) models are more generalizable to oth er domains. On the other hand, there lacks of theoretical understanding of such phenomenon and it is not clear whether there are counterexamples. In this paper, we aim to provide sufficient conditions for this phenomenon considering different factors that could affect both, such as the norm of last layer norm, Jacobian norm, and data augmentations (DA). In particular, we propose a general theoretical framework indicating factors that can be reformed as a function class regularization process, which could lead to the improvement of domain generalization. Our analysis, for the first time, shows that ``robustness" is actually not the causation for domain generalization; rather, robustness induced by adversarial training is a by-product of such function class regularization. We then discuss in details about different properties of DA and we prove that under certain conditions, DA can be viewed as regularization and therefore improve generalization. We conduct extensive experiments to verify our theoretical findings, and show several counterexamples where robustness and generalization are negatively correlated when the sufficient conditions are not satisfied.

****

Parallel Deep Neural Networks Have Zero Duality Gap

Yifei Wang,Tolga Ergen,Mert Pilanci

Training deep neural networks is a well-known highly non-convex problem. In recent works, it is shown that there is no duality gap for regularized two-layer neural networks with ReLU activation, which enables global optimization via convex programs. For multi-layer linear networks with vector outputs, we formulate convex dual problems and demonstrate that the duality gap is non-zero for depth three and deeper networks. However, by modifying the deep networks to more powerful parallel architectures, we show that the duality gap is exactly zero. Therefore, strong convex duality holds, and hence there exist equivalent convex programs that enable training deep networks to global optimality. We also demonstrate that the weight decay regularization in the parameters explicitly encourages low-rank solutions via closed-form expressions. For three-layer non-parallel ReLU networks, we show that strong duality holds for rank-1 data matrices, however, the duality gap is non-zero for whitened data matrices. Similarly, by transforming the neural network architecture into a corresponding parallel version, the duality gap vanishes.

****

Minibatch vs Local SGD with Shuffling: Tight Convergence Bounds and Beyond

Chulhee Yun,Shashank Rajput,Suvrit Sra

In distributed learning, local SGD (also known as federated averaging) and its simple baseline minibatch SGD are widely studied optimization methods. Most existing analyses of these methods assume independent and unbiased gradient estimates obtained via with-replacement sampling. In contrast, we study shuffling-based variants: minibatch and local Random Reshuffling, which draw stochastic gradients without replacement and are thus closer to practice. For smooth functions satisfying the Polyak-■ojasiewicz condition, we obtain convergence bounds (in the large epoch regime) which show that these shuffling-based variants converge faster than their with-replacement counterparts. Moreover, we prove matching lower bounds showing that our convergence analysis is tight. Finally, we propose an algorithmic modification called synchronized shuffling that leads to convergence rates faster than our lower bounds in near-homogeneous settings.

****

Learning to Map for Active Semantic Goal Navigation

Georgios Georgakis,Bernadette Bucher,Karl Schmeckpeper,Siddharth Singh,Kostas Daniilidis

We consider the problem of object goal navigation in unseen environments. Solving this problem requires learning of contextual semantic priors, a challenging endeavour given the spatial and semantic variability of indoor environments. Current methods learn to implicitly encode these priors through goal-oriented navigation policy functions operating on spatial representations that are limited to the agent's observable areas. In this work, we propose a novel framework that actively learns to generate semantic maps outside the field of view of the agent and leverages the uncertainty over the semantic classes in the unobserved areas to decide on long term goals. We demonstrate that through this spatial prediction strategy, we are able to learn semantic priors in scenes that can be leveraged in unknown environments. Additionally, we show how different objectives can be defined by balancing exploration with exploitation during searching for semantic targets. Our method is validated in the visually realistic environments of the Matterport3D dataset and show improved results on object goal navigation over competitive baselines.

****

Benchmarking the Spectrum of Agent Capabilities

Danijar Hafner

Evaluating the general abilities of intelligent agents requires complex simulation environments. Existing benchmarks typically evaluate only one narrow task per environment, requiring researchers to perform expensive training runs on many different environments. We introduce Crafter, an open world survival game with visual inputs that evaluates a wide range of general abilities within a single environment. Agents either learn from the provided reward signal or through intrinsic objectives and are evaluated by semantically meaningful achievements that can be unlocked during each episode, such as discovering resources and crafting tools. Consistently unlocking all achievements requires strong generalization, deep exploration, and long-term reasoning. We experimentally verify that Crafter is of appropriate difficulty to drive future research and provide baselines scores of reward agents and unsupervised agents. Furthermore, we observe sophisticated behaviors emerging from maximizing the reward signal, such as building tunnel systems, bridges, houses, and plantations. We hope that Crafter will accelerate research progress by quickly evaluating a wide spectrum of abilities.

****

Mind the Gap: Domain Gap Control for Single Shot Domain Adaptation for Generative Adversarial Networks

Peihao Zhu,Rameen Abdal,John Femiani,Peter Wonka

We present a new method for one shot domain adaptation. The input to our method is trained GAN that can produce images in domain A and a single reference image I_B from domain B. The proposed algorithm can translate any output of the trained GAN from domain A to domain B. There are two main advantages of our method compared to the current state of the art: First, our solution achieves higher visual quality, e.g. by noticeably reducing overfitting. Second, our solution allows for more degrees of freedom to control the domain gap, i.e. what aspects of image I_B are used to define the domain B. Technically, we realize the new method by building on a pre-trained StyleGAN generator as GAN and a pre-trained CLIP model for representing the domain gap. We propose several new regularizers for controlling the domain gap to optimize the weights of the pre-trained StyleGAN generator to output images in domain B instead of domain A. The regularizers prevent the optimization from taking on too many attributes of the single reference image. Our results show significant visual improvements over the state of the art as well as multiple applications that highlight improved control.

****

The Hidden Convex Optimization Landscape of Regularized Two-Layer ReLU Networks: an Exact Characterization of Optimal Solutions

Yifei Wang,Jonathan Lacotte,Mert Pilanci

We prove that finding all globally optimal two-layer ReLU neural networks can be performed by solving a convex optimization program with cone constraints. Our analysis is novel, characterizes all optimal solutions, and does not leverage duality-based analysis which was recently used to lift neural network training into convex spaces. Given the set of solutions of our convex optimization program, we show how to construct exactly the entire set of optimal neural networks. We provide a detailed characterization of this optimal set and its invariant transformations. As additional consequences of our convex perspective, (i) we establish that Clarke stationary points found by stochastic gradient descent correspond to the global optimum of a subsampled convex problem (ii) we provide a polynomial-time algorithm for checking if a neural network is a global minimum of the training loss (iii) we provide an explicit construction of a continuous path between any neural network and the global minimum of its sublevel set and (iv) characterize the minimal size of the hidden layer so that the neural network optimization landscape has no spurious valleys.

Overall, we provide a rich framework for studying the landscape of neural network training loss through convexity.

****

FSL: Federated Supermask Learning

Hamid Mozaffari,Virat Shejwalkar,Amir Houmansadr

Federated learning (FL) allows multiple clients with (private) data to collaboratively train a common machine learning model without sharing their private training data. In-the-wild deployment of FL faces two major hurdles: robustness to poisoning attacks and communication efficiency. To address these concurrently, we propose Federated Supermask Learning (FSL). FSL server trains a global subnetwork within a randomly initialized neural network by aggregating local subnetworks of all collaborating clients. FSL clients share local subnetworks in the form of rankings of network edges; more useful edges have higher ranks. By sharing integer rankings, instead of float weights, FSL restricts the space available to craft effective poisoning updates, and by sharing subnetworks, FSL reduces the communication cost of training. We show theoretically and empirically that FSL is robust by design and also significantly communication efficient; all this without compromising clients' privacy. Our experiments demonstrate the superiority of FSL in real-world FL settings; in particular, (1) FSL achieves similar performances as state-of-the-art FedAvg with significantly lower communication costs: for CIFAR10, FSL achieves same performance as Federated Averaging while reducing communication cost by $\sim35\%$. (2) FSL is substantially more robust to poisoning attacks than state-of-the-art robust aggregation algorithms.

****

Improving and Assessing Anomaly Detectors for Large-Scale Settings

Dan Hendrycks,Steven Basart,Mantas Mazeika,Andy Zou,Joseph Kwon,Mohammadreza Mostajabi,Jacob Steinhardt

Detecting out-of-distribution examples is important for safety-critical machine learning applications such as detecting novel biological phenomena and self-driving cars. However, existing research mainly focuses on simple small-scale settings. To set the stage for more realistic out-of-distribution detection, we depart from small-scale settings and explore large-scale multiclass and multi-label settings with high-resolution images and thousands of classes. To make future work in real-world settings possible, we create new benchmarks for three large-scale settings. To test ImageNet multiclass anomaly detectors, we introduce a new dataset of anomalous species. We leverage ImageNet-22K to evaluate PASCAL VOC and COCO multilabel anomaly detectors. Third, we introduce a new benchmark for anomaly segmentation by introducing a segmentation benchmark with road anomalies. We conduct extensive experiments in these more realistic settings for out-of-distribution detection and find that a surprisingly simple detector based on the maximum logit outperforms prior methods in all the large-scale multi-class, multi-label, and segmentation tasks, establishing a simple new baseline for future work.

****

On Evaluation Metrics for Graph Generative Models

Rylee Thompson,Boris Knyazev,Elahe Ghalebi,Jungtaek Kim,Graham W. Taylor

In image generation, generative models can be evaluated naturally by visually inspecting model outputs. However, this is not always the case for graph generative models (GGMs), making their evaluation challenging. Currently, the standard process for evaluating GGMs suffers from three critical limitations: i) it does not produce a single score which makes model selection challenging, ii) in many cases it fails to consider underlying edge and node features, and iii) it is prohibitively slow to perform. In this work, we mitigate these issues by searching for \emph{scalar, domain-agnostic, and scalable metrics} for evaluating and ranking GGMs. To this end, we study existing GGM metrics and neural-network-based metrics emerging from generative models of images that use embeddings extracted from a task-specific network. Motivated by the power of Graph Neural Networks (GNNs) to extract meaningful graph representations \emph{without any training}, we int roduce several metrics based on the features extracted by an untrained random GNN. We design experiments to thoroughly test and objectively score metrics on their ability to measure the diversity and fidelity of generated graphs, as well as their sample and computational efficiency. Depending on the quantity of samples, we recommend one of two metrics from our collection of random-GNN-based metrics. We show these two metrics to be more expressive than pre-existing and alternative random-GNN-based metrics using our objective scoring. While we focus on applying these metrics to GGM evaluation, in practice this enables the ability to easily compute the dissimilarity between any two sets of graphs \emph{regardless of domain}. Our code is released at: https://github.com/uoguelph-mlrg/GGM-metrics.

****

Value-aware transformers for 1.5d data
James F Cann,Timothy J Roberts,Amy R Tso,Amy Nelson,Parashkev Nachev
Sparse sequential highly-multivariate data of the form characteristic of hospital in-patient investigation and treatment poses a considerable challenge for representation learning. Such data is neither faithfully reducible to 1d nor dense enough to constitute multivariate series. Conventional models compromise their data by requiring these forms at the point of input. Building on contemporary sequence-modelling architectures we design a value-aware transformer, prompting a reconceptualisation of our data as 1.5-dimensional: a token-value form both respecting its sequential nature and augmenting it with a quantifier. Experiments focused on sequential in-patient laboratory data up to 48hrs after hospital admission show that the value-aware transformer performs favourably versus competitive baselines on in-hospital mortality and length-of-stay prediction within the MIMIC-III dataset.

****

Selective Ensembles for Consistent Predictions
Emily Black,Klas Leino,Matt Fredrikson
Recent work has shown that models trained to the same objective, and which achieve similar measures of accuracy on consistent test data, may nonetheless behave very differently on individual predictions. This inconsistency is undesirable in high-stakes contexts, such as medical diagnosis and finance. We show that this duplicitous behavior extends beyond predictions to feature attributions, which may likewise have negative implications for the intelligibility of a model, and one's ability to find recourse for subjects. We then introduce selective ensembles to mitigate such inconsistencies by applying hypothesis testing to the predictions of a set of models trained using randomly-selected starting conditions; importantly, selective ensembles can abstain in cases where a consistent outcome cannot be achieved up to a specified confidence level. We prove that that prediction disagreement between selective ensembles is bounded, and empirically demonstrate that selective ensembles achieve consistent predictions and feature attributions while maintaining low abstention rates. On several benchmark datasets, selective ensembles reach zero inconsistently predicted points, with abstention rates as low as 1.5%.

****

Graph Condensation for Graph Neural Networks
Wei Jin,Lingxiao Zhao,Shichang Zhang,Yozen Liu,Jiliang Tang,Neil Shah
Given the prevalence of large-scale graphs in real-world applications, the storage and time for training neural models have raised increasing concerns. To alleviate the concerns, we propose and study the problem of graph condensation for graph neural networks (GNNs). Specifically, we aim to condense the large, original graph into a small, synthetic and highly-informative graph, such that GNNs trained on the small graph and large graph have comparable performance. We approach the condensation problem by imitating the GNN training trajectory on the original graph through the optimization of a gradient matching loss and design a strategy to condense node futures and structural information simultaneously. Extensive experiments have demonstrated the effectiveness of the proposed framework in condensing different graph datasets into informative smaller graphs. In particular, we are able to approximate the original test accuracy by 95.3\% on Reddit, 9

9.8\% on Flickr and 99.0\% on Citeseer, while reducing their graph size by more than 99.9\%, and the condensed graphs can be used to train various GNN architectures.

****

DIVA: Dataset Derivative of a Learning Task

Yonatan Dukler,Alessandro Achille,Giovanni Paolini,Avinash Ravichandran,Marzia Polito,Stefano Soatto

We present a method to compute the derivative of a learning task with respect to a dataset. A learning task is a function from a training set to the validation error, which can be represented by a trained deep neural network (DNN). The ``dataset derivative'' is a linear operator, computed around the trained model, that informs how perturbations of the weight of each training sample affect the validation error, usually computed on a separate validation dataset. Our method, DIVA (Differentiable Validation) hinges on a closed-form differentiable expression of the leave-one-out cross-validation error around a pre-trained DNN. Such expression constitutes the dataset derivative. DIVA could be used for dataset auto-curation, for example removing samples with faulty annotations, augmenting a dataset with additional relevant samples, or rebalancing. More generally, DIVA can be used to optimize the dataset, along with the parameters of the model, as part of the training process without the need for a separate validation dataset, unlike bi-level optimization methods customary in AutoML. To illustrate the flexibility of DIVA, we report experiments on sample auto-curation tasks such as outlier rejection, dataset extension, and automatic aggregation of multi-modal data.

****

Poisoned classifiers are not only backdoored, they are fundamentally broken

Mingjie Sun,Siddhant Agarwal,J Zico Kolter

Under a commonly-studied backdoor poisoning attack against classification models, an attacker adds a small trigger to a subset of the training data, such that the presence of this trigger at test time causes the classifier to always predict some target class. It is often implicitly assumed that the poisoned classifier is vulnerable exclusively to the adversary who possesses the trigger. In this paper, we show empirically that this view of backdoored classifiers is incorrect. We describe a new threat model for poisoned classifier, where one without knowledge of the original trigger, would want to control the poisoned classifier. Under this threat model, we propose a test-time, human-in-the-loop attack method to generate multiple effective alternative triggers without access to the initial backdoor and the training data. We construct these alternative triggers by first generating adversarial examples for a smoothed version of the classifier, created with a procedure called Denoised Smoothing, and then extracting colors or cropped portions of smoothed adversarial images with human interaction. We demonstrate the effectiveness of our attack through extensive experiments on high-resolution datasets: ImageNet and TrojAI. We also compare our approach to previous work on modeling trigger distributions and find that our method are more scalable and efficient in generating effective triggers. Last, we include a user study which demonstrates that our method allows users to easily determine the existence of such backdoors in existing poisoned classifiers. Thus, we argue that there is no such thing as a secret backdoor in poisoned classifiers: poisoning a classifier invites attacks not just by the party that possesses the trigger, but from anyone with access to the classifier.

****

Towards General Function Approximation in Zero-Sum Markov Games

Baihe Huang,Jason D. Lee,Zhaoran Wang,Zhuoran Yang

This paper considers two-player zero-sum finite-horizon Markov games with simultaneous moves. The study focuses on the challenging settings where the value function or the model is parameterized by general function classes. Provably efficient algorithms for both decoupled and coordinated settings are developed. In the decoupled setting where the agent controls a single player and plays against an arbitrary opponent, we propose a new model-free algorithm. The sample complexity is governed by the Minimax Eluder dimension—a new dimension of the function class in Markov games. As a special case, this method improves the state-of-the-art algorithm by a $\sqrt{d}$ factor in the regret when the reward function and transition kernel are parameterized with d-dimensional linear features. In the coordinated setting where both players are controlled by the agent, we propose a model-based algorithm and a model-free algorithm. In the model-based algorithm, we prove that sample complexity can be bounded by a generalization of Witness rank to Markov games. The model-free algorithm enjoys a $\sqrt{K}$-regret upper bound where K is the number of episodes. Our algorithms are based on new techniques of alternate optimism

****

Spectral Multiplicity Entails Sample-wise Multiple Descent

Lin Chen,Song Mei

In this paper, we study the generalization risk of ridge and ridgeless linear regression. We assume that the data features follow a multivariate normal distribution and that the spectrum of the covariance matrix consists of a given set of eigenvalues of proportionally growing multiplicity. We characterize the limiting bias and variance when the dimension and the number of training samples tend to infinity proportionally. Exact formulae for the bias and variance are derived using the random matrix theory and convex Gaussian min-max theorem. Based on these formulae, we study the sample-wise multiple descent phenomenon of the generalization risk curve, i.e., with more data, the generalization risk can be non-monotone, and specifically, can increase and then decrease multiple times with more training data samples. We prove that sample-wise multiple descent occurs when the spectrum of the covariance matrix is highly ill-conditioned. We also present numerical results to confirm the values of the bias and variance predicted by our theory and illustrate the multiple descent of the generalization risk curve. Moreover, we theoretically show that the ridge estimator with optimal regularization can result in a monotone generalization risk curve and thereby eliminate multiple descent under some assumptions.

****

Exposing the Implicit Energy Networks behind Masked Language Models via Metropolis--Hastings

Kartik Goyal,Chris Dyer,Taylor Berg-Kirkpatrick

While recent work has shown that scores from models trained by the ubiquitous masked language modeling (MLM) objective effectively discriminate probable from improbable sequences, it is still an open question if these MLMs specify a principled probability distribution over the space of possible sequences. In this paper, we interpret MLMs as energy-based sequence models and propose two energy parametrizations derivable from the trained MLMs. In order to draw samples correctly from these models, we develop a tractable sampling scheme based on the Metropolis--Hastings Monte Carlo algorithm. In our approach, samples are proposed from the same masked conditionals used for training the masked language models, and they are accepted or rejected based on their energy values according to the target distribution. We validate the effectiveness of the proposed parametrizations by exploring the quality of samples drawn from these energy-based models for both open-ended unconditional generation and a conditional generation task of machine translation. We theoretically and empirically justify our sampling algorithm by showing that the masked conditionals on their own do not yield a Markov chain whose stationary distribution is that of our target distribution, and our approach generates higher quality samples than other recently proposed undirected generation approaches (Wang et al., 2019, Ghazvininejad et al., 2019).

****

The guide and the explorer: smart agents for resource-limited iterated batch reinforcement learning

Albert Thomas,Balázs Kégl,Othman Gaizi,Gabriel Hurtado

Iterated batch reinforcement learning (RL) is a growing subfield fueled by the demand from systems engineers for intelligent control solutions that they can app ly within their technical and organizational constraints. Model-based RL (MBRL) suits this scenario well for its sample efficiency and modularity. Recent MBRL techniques combine efficient neural system models with classical planning (like model predictive control; MPC). In this paper we add two components to this classical setup. The first is a Dyna-style policy learned on the system model using model-free techniques. We call it the guide since it guides the planner. The second component is the explorer, a strategy to expand the limited knowledge of the guide during planning. Through a rigorous ablation study we show that exploration is crucial for optimal performance. We apply this approach with a DQN guide and a heating explorer to improve the state of the art of the resource-limited Acrobot benchmark system by about 10%.

****

ClimateGAN: Raising Climate Change Awareness by Generating Images of Floods

Victor Schmidt,Alexandra Luccioni,Mélisande Teng,Tianyu Zhang,Alexia Reynaud,Sunand Raghupathi,Gautier Cosne,Adrien Juraver,Vahe Vardanyan,Alex Hernández-García,Yoshua Bengio

Climate change is a major threat to humanity and the actions required to prevent its catastrophic consequences include changes in both policy-making and individual behaviour. However, taking action requires understanding its seemingly abstract and distant consequences. Projecting the potential impacts of extreme climate events such as flooding in familiar places can help make the impacts of climate change more concrete and encourage action. As part of a larger initiative to build a website (https://thisclimatedoesnotexist.com) that projects extreme climate events onto user-chosen photos, we present our solution to simulate photo-realistic floods on authentic images. To address this complex task in the absence of suitable data, we propose ClimateGAN, a model that leverages both simulated and real data through unsupervised domain adaptation and conditional image generation. In this paper, we describe the details of our framework, thoroughly evaluate the main components of our architecture and demonstrate that our model is capable of robustly generating photo-realistic flooding on street images.

****

FedLite: A Scalable Approach for Federated Learning on Resource-constrained Clients

Jianyu Wang,Hang Qi,Ankit Singh Rawat,Sashank J. Reddi,Sagar M. Waghmare,Felix Yu,Gauri Joshi

In classical federated learning, the clients contribute to the overall training by communicating local updates for the underlying model on their private data to a coordinating server. However, updating and communicating the entire model becomes prohibitively expensive when resource-constrained clients collectively aim to train a large machine learning model. Split learning provides a natural solution in such a setting, where only a (small) part of the model is stored and trained on clients while the remaining (large) part of the model only stays at the servers. Unfortunately, the model partitioning employed in split learning significantly increases the communication cost compared to the classical federated learning algorithms. This paper addresses this issue by proposing an end-to-end training framework that relies on a novel vector quantization scheme accompanied by a gradient correction method to reduce the additional communication cost associated with split learning. An extensive empirical evaluation on standard image and text benchmarks shows that the proposed method can achieve up to $490\times$ communication cost reduction with minimal drop in accuracy, and enables a desirable performance vs. communication trade-off.

****

A Comparison of Hamming Errors of Representative Variable Selection Methods

Tracy Ke,Longlin Wang

Lasso is a celebrated method for variable selection in linear models, but it faces challenges when the covariates are moderately or strongly correlated. This motivates alternative approaches such as using a non-convex penalty, adding a ridge regularization, or conducting a post-Lasso thresholding. In this paper, we compare Lasso with 5 other methods: Elastic net, SCAD, forward selection, thresholded Lasso, and forward backward selection. We measure their performances theoreti cally by the expected Hamming error, assuming that the regression coefficients are ${\it iid}$ drawn from a two-point mixture and that the Gram matrix is block-wise diagonal. By deriving the rates of convergence of Hamming errors and the phase diagrams, we obtain useful conclusions about the pros and cons of different methods.

****

On Convergence of Federated Averaging Langevin Dynamics

Wei Deng,Yian Ma,Zhao Song,Qian Zhang,Guang Lin

We propose a federated averaging Langevin algorithm (FA-LD) for uncertainty quantification and mean predictions with distributed clients. In particular, we generalize beyond normal posterior distributions and consider a general class of models. We develop theoretical guarantees for FA-LD for strongly log-concave distributions with non-i.i.d data and study how the injected noise and the stochastic-gradient noise, the heterogeneity of data, and the varying learning rates affect the convergence. Such an analysis sheds light on the optimal choice of local updates to minimize communication cost. Important to our approach is that the communication efficiency does not deteriorate with the injected noise in the Langevin algorithms. In addition, we examine in our FA-LD algorithm both independent and correlated noise used over different clients. We observe that there is also a trade-off between federation and communication cost there. As local devices may become inactive in the federated network, we also show convergence results based on different averaging schemes where only partial device updates are available.

****

Latent Variable Sequential Set Transformers for Joint Multi-Agent Motion Prediction

Roger Girgis,Florian Golemo,Felipe Codevilla,Martin Weiss,Jim Aldon D'Souza,Samira Ebrahimi Kahou,Felix Heide,Christopher Pal

Robust multi-agent trajectory prediction is essential for the safe control of robotic systems. A major challenge is to efficiently learn a representation that approximates the true joint distribution of contextual, social, and temporal information to enable planning. We propose Latent Variable Sequential Set Transformers which are encoder-decoder architectures that generate scene-consistent multi-agent trajectories. We refer to these architectures as “AutoBots”. The encoder is a stack of interleaved temporal and social multi-head self-attention (MHSA) modules which alternately perform equivariant processing across the temporal and social dimensions. The decoder employs learnable seed parameters in combination with temporal and social MHSA modules allowing it to perform inference over the entire future scene in a single forward pass efficiently. AutoBots can produce either the trajectory of one ego-agent or a distribution over the future trajectories for all agents in the scene. For the single-agent prediction case, our model achieves top results on the global nuScenes vehicle motion prediction leaderboard, and produces strong results on the Argoverse vehicle prediction challenge. In the multi-agent setting, we evaluate on the synthetic partition of TrajNet++ dataset to showcase the model’s socially-consistent predictions. We also demonstrate our model on general sequences of sets and provide illustrative experiments modelling the sequential structure of the multiple strokes that make up symbols in the Omniglot data. A distinguishing feature of AutoBots is that all models are trainable on a
single desktop GPU (1080 Ti) in under 48h.

****

A Program to Build E(N)-Equivariant Steerable CNNs

Gabriele Cesa,Leon Lang,Maurice Weiler

Equivariance is becoming an increasingly popular design choice to build data efficient neural networks by exploiting prior knowledge about the symmetries of the problem at hand. Euclidean steerable CNNs are one of the most common classes of equivariant networks. While the constraints these architectures need to satisfy are understood, existing approaches are tailored to specific (classes of) groups. No generally applicable method that is practical for implementation has been described so far. In this work, we generalize the Wigner-Eckart theorem proposed in Lang & Weiler (2020), which characterizes general G-steerable kernel space s for compact groups G over their homogeneous spaces, to arbitrary G-spaces. This enables us to directly parameterize filters in terms of a band-limited basis on the whole space rather than on G's orbits, but also to easily implement steerable CNNs equivariant to a large number of groups. To demonstrate its generality, we instantiate our method on a variety of isometry groups acting on the Euclidean space $\mathbb{R}^3$. Our framework allows us to build $E(3)$ and $SE(3)$-steerable CNNs like previous works, but also CNNs with arbitrary $G\leq O(3)$-steerable kernels. For example, we build 3D CNNs equivariant to the symmetries of platonic solids or choose $G=SO(2)$ when working with 3D data having only azimuthal symmetries. We compare these models on 3D shapes and molecular datasets, observing improved performance by matching the model's symmetries to the ones of the data.

****

Minimax Optimization with Smooth Algorithmic Adversaries

Tanner Fiez,Chi Jin,Praneeth Netrapalli,Lillian J Ratliff

This paper considers minimax optimization $\min_x \max_y f(x, y)$ in the challenging setting where f can be both nonconvex in x and nonconcave in y. Though such optimization problems arise in many machine learning paradigms including training generative adversarial networks (GANs) and adversarially robust models, from a theoretical point of view, two fundamental issues remain: (i) the absence of simple and efficiently computable optimality notions, and (ii) cyclic or diverging behavior of existing algorithms. This paper proposes a new theoretical framework for nonconvex-nonconcave minimax optimization that addresses both of the above issues. The starting point of this paper is the observation that, under a computational budget, the max-player can not fully maximize $f(x,\cdot)$ since nonconcave maximization is NP-hard in general. So, we propose a new framework, and a corresponding algorithm, for the min-player to play against \emph{smooth algorithms} deployed by the adversary (i.e., the max-player) instead of against full maximization. Our algorithm is guaranteed to make monotonic progress (thus having no limit cycles or diverging behavior), and to find an appropriate ``stationary point'' in a polynomial number of iterations. Our framework covers practically relevant settings where the smooth algorithms deployed by the adversary are multi-step stochastic gradient ascent, and its accelerated version. We further present experimental results that confirm our theoretical findings and demonstrate the effectiveness of the proposed approach in practice on simple, conceptual settings.

****

Attention: Self-Expression Is All You Need

Rene Vidal

Transformer models have achieved significant improvements in performance for various learning tasks in natural language processing and computer vision. Much of their success is attributed to the use of attention layers that capture long-range interactions among data tokens (such as words and image patches) via attention coefficients that are global and adapted to the input data at test time. In this paper we study the principles behind attention and its connections with prior art. Specifically, we show that attention builds upon a long history of prior work on manifold learning and image processing, including methods such as kernel-based regression, non-local means, locally linear embedding, subspace clustering and sparse coding. Notably, we show that self-attention is closely related to the notion of self-expressiveness in subspace clustering, wherein data points to be clustered are expressed as linear combinations of other points with global coefficients that are adapted to the data and capture long-range interactions among data points. We also show that heuristics in sparse self-attention can be studied in a more principled manner using prior literature on sparse coding and sparse subspace clustering. We thus conclude that the key innovations of attention mechanisms relative to prior art are the use of many learnable parameters, and multiple heads and layers.

****

Revealing the Incentive to Cause Distributional Shift

David Krueger,Tegan Maharaj,Jan Leike

Decisions made by machine learning systems have increasing influence on the world, yet it is common for machine learning algorithms to assume that no such influence exists. An example is the use of the i.i.d. assumption in content recommendation: In fact, the (choice of) content displayed can change users’ perceptions and preferences, or even drive them away, causing a shift in the distribution of users. We introduce the term auto-induced distributional shift (ADS) to describe the phenomenon of an algorithm causing change in the distribution of its own inputs. Leveraging ADS can be a means of increasing performance. But this is not always desirable, since performance metrics often underspecify what type of behaviour is desirable. When real-world conditions violate assumptions (such as i.i.d. data), this underspecification can result in unexpected behaviour. To diagnose such issues, we introduce the approach of unit tests for incentives: simple environments designed to show whether an algorithm will hide or reveal incentives to achieve performance via certain means (in our case, via ADS). We use these unit tests to demonstrate that changes to the learning algorithm (e.g. introducing meta-learning) can cause previously hidden incentives to be revealed, resulting in qualitatively different behaviour despite no change in performance metric. We further introduce a toy environment for modelling real-world issues with ADS in content recommendation, where we demonstrate that strong meta-learners achieve gains in performance via ADS. These experiments confirm that the unit tests work – an algorithm’s failure of the unit test correctly diagnoses its propensity to reveal incentives for ADS.

****

Provably Filtering Exogenous Distractors using Multistep Inverse Dynamics

Yonathan Efroni,Dipendra Misra,Akshay Krishnamurthy,Alekh Agarwal,John Langford

Many real-world applications of reinforcement learning (RL) require the agent to deal with high-dimensional observations such as those generated from a megapixel camera. Prior work has addressed such problems with representation learning, through which the agent can provably extract endogenous, latent state information from raw observations and subsequently plan efficiently. However, such approaches can fail in the presence of temporally correlated noise in the observations, a phenomenon that is common in practice. We initiate the formal study of latent state discovery in the presence of such exogenous noise sources by proposing a new model, the Exogenous Block MDP (EX-BMDP), for rich observation RL. We start by establishing several negative results, by highlighting failure cases of prior representation learning based approaches. Then, we introduce the Predictive Path Elimination (PPE) algorithm, that learns a generalization of inverse dynamics and is provably sample and computationally efficient in EX-BMDPs when the endogenous state dynamics are near deterministic. The sample complexity of PPE depends polynomially on the size of the latent endogenous state space while not directly depending on the size of the observation space, nor the exogenous state space. We provide experiments on challenging exploration problems which show that our approach works empirically.

****

Language Models as Zero-Shot Planners: Extracting Actionable Knowledge for Embodied Agents

Wenlong Huang,Pieter Abbeel,Deepak Pathak,Igor Mordatch

Can world knowledge learned by large language models (LLMs) be used to act in interactive environments? In this paper, we investigate the possibility of grounding high-level tasks, expressed in natural language (i.e. "make breakfast"), to a fixed set of actionable steps (i.e. "open fridge"). While prior work focused on learning from explicit step-by-step examples of how to act, we surprisingly find that if pre-trained LMs are large enough and prompted appropriately, they can effectively decompose high-level tasks into low-level plans without any further training. However, the plans produced naively by LLMs often cannot map precisely to admissible actions. We propose a procedure that conditions on existing demonstrations and semantically translates the plans to admissible actions. Our evaluation in the recent VirtualHome environment shows that the resulting method substantially improves executability over the LLM baseline. The conducted human evaluation reveals a trade-off between executability and correctness but shows a pro mising sign towards extracting actionable knowledge from language models. Videos at https://sites.google.com/view/language-model-as-planner

****

Ensemble Kalman Filter (EnKF) for Reinforcement Learning (RL)

Anant A Joshi,Amirhossein Taghvaei,Prashant G Mehta

This paper is concerned with representing and learning the optimal control law for the linear quadratic Gaussian (LQG) optimal control problem. In recent years, there is a growing interest in re-visiting this classical problem, in part due to the successes of reinforcement learning (RL). The main question of this body of research (and also of our paper) is to approximate the optimal control law without explicitly solving the Riccati equation. For this purpose, a novel simulation-based algorithm, namely an ensemble Kalman filter (EnKF), is introduced in this paper. The algorithm is used to obtain formulae for optimal control, expressed entirely in terms of the EnKF particles. For the general partially observed LQG problem, the proposed EnKF is combined with a standard EnKF (for the estimation problem) to obtain the optimal control input based on the use of the separation principle. The theoretical results and algorithms are illustrated with numerical experiments.

****

On the exploitative behavior of adversarial training against adversarial attacks

Ali Rahmati,Seyed-Mohsen Moosavi-Dezfooli,Huaiyu Dai

Adversarial attacks have been developed as intentionally designed perturbations added to the inputs in order to fool deep neural network classifiers. Adversarial training has been shown to be an effective approach to improving the robustness of the classifiers against such attacks especially in the white-box setting. In this work, we demonstrate that some geometric consequences of adversarial training on the decision boundary of deep networks give an edge to certain types of black-box attacks. In particular, we introduce a highly parallelizable black-box attack against the classifiers equipped with an ℓ_2 norm similarity detector, which exploits the low mean curvature of the decision boundary. We use this black-box attack to demonstrate that adversarially-trained networks might be easier to fool in certain scenarios. Moreover, we define a metric called robustness gain to show that while adversarial training is an effective method to improve the robustness in the white-box attack setting, it may not provide such a good robustness gain against the more realistic decision-based black-box attacks.

****

Teacher's pet: understanding and mitigating biases in distillation

Michal Lukasik,Srinadh Bhojanapalli,Aditya Krishna Menon,Sanjiv Kumar

Knowledge distillation is widely used as a means of improving the performance of a relatively simple “student” model using the predictions from a complex “teacher” model. Several works have shown that distillation significantly boosts the student’s overall performance; however, are these gains uniform across all data sub-groups? In this paper, we show that distillation can harm performance on certain subgroups, e.g., classes with few associated samples, compared to the vanilla student trained using the one-hot labels. We trace this behavior to errors made by the teacher distribution being transferred to and amplified by the student model. To mitigate this problem, we present techniques which soften the teacher influence for subgroups where it is less reliable. Experiments on several image classification benchmarks show that these modifications of distillation maintain boost in overall accuracy, while additionally ensuring improvement in subgroup performance.

****

Zero-Shot Reward Specification via Grounded Natural Language

Parsa Mahmoudieh,Sayna Ebrahimi,Deepak Pathak,Trevor Darrell

Reward signals in reinforcement learning can be expensive signals in many tasks and often require access to direct state. The alternative to reward signals are usually demonstrations or goal images which can be labor intensive to collect. Goal text description is a low effort way of communicating the desired task. Goal text conditioned policies so far though have been trained with reward signals that have access to state or labelled expert demonstrations. We devise a model th at leverages CLIP to ground objects in a scene described by the goal text paired with spatial relationship rules to provide an off-the-shelf reward signal on only raw pixels to learn a set of robotic manipulation tasks. We distill the policies learned with this reward signal on several tasks to produce one goal text conditioned policy.

****

On Distributed Adaptive Optimization with Gradient Compression
Xiaoyun Li,Belhal Karimi,Ping Li

We study COMP-AMS, a distributed optimization framework based on gradient averaging and adaptive AMSGrad algorithm. Gradient compression with error feedback is applied to reduce the communication cost in the gradient transmission process. Our convergence analysis of COMP-AMS shows that such compressed gradient averaging strategy yields same convergence rate as standard AMSGrad, and also exhibits the linear speedup effect w.r.t. the number of local workers. Compared with recently proposed protocols on distributed adaptive methods, COMP-AMS is simple and convenient. Numerical experiments are conducted to justify the theoretical findings, and demonstrate that the proposed method can achieve same test accuracy as the full-gradient AMSGrad with substantial communication savings. With its simplicity and efficiency, COMP-AMS can serve as a useful distributed training framework for adaptive methods.

****

Learning Symmetric Representations for Equivariant World Models
Jung Yeon Park,Ondrej Biza,Linfeng Zhao,Jan-Willem van de Meent,Robin Walters

Encoding known symmetries into world models can improve generalization. However, identifying how latent symmetries manifest in the input space can be difficult. As an example, rotations of objects are equivariant with respect to their orientation, but extracting this orientation from an image is difficult in absence of supervision. In this paper, we use equivariant transition models as an inductive bias to learn symmetric latent representations in a self-supervised manner. This allows us to train non-equivariant networks to encode input data, for which the underlying symmetry may be non-obvious, into a latent space where symmetries may be used to reason about outcomes of actions in a data-efficient manner. Our method is agnostic to the type of latent symmetry; we demonstrate its usefulness over $C_4 \times S_5$ using G-convolutions and GNNs, over $D_4 \ltimes (\mathbb{R}^2,+)$ using $E(2)$-steerable CNNs, and over $\mathrm{SO}(3)$ using tensor field networks. In all three cases, we demonstrate improvements relative to both fully-equivariant and non-equivariant baselines.

****

Neural Networks Playing Dough: Investigating Deep Cognition With a Gradient-Based Adversarial Attack
Sahar Niknam

Discovering adversarial examples has shaken our trust in the reliability of deep learning. Even though brilliant works have been devoted to understanding and fixing this vulnerability, fundamental questions (e.g. the mysterious generalization of adversarial examples across models and training sets) remain unanswered. This paper tests the hypothesis that it is not the neural networks failing in learning that causes adversarial vulnerability, but their different perception of the presented data. And therefore, adversarial examples should be semantic-sensitive signals which can provide us with an exceptional opening to understanding neural network learning. To investigate this hypothesis, I performed a gradient-based attack on fully connected feed-forward and convolutional neural networks, instructing them to minimally evolve controlled inputs into adversarial examples for all the classes of the MNIST and Fashion-MNIST datasets. Then I abstracted adversarial perturbations from these examples. The perturbations unveiled vivid and recurring visual structures, unique to each class and persistent over parameters of abstraction methods, model architectures, and training configurations. Furthermore, these patterns proved to be explainable and derivable from the corresponding dataset. This finding explains the generalizability of adversarial examples by, semantically, tying them to the datasets. In conclusion, this experiment not only resists interpretation of adversarial examples as deep learning failure

but on the contrary, demystifies them in the form of supporting evidence for the authentic learning capacity of neural networks.
****
Interrogating Paradigms in Self-supervised Graph Representation Learning
Puja Trivedi,Mark Heimann,Danai Koutra,Jayaraman J. Thiagarajan
Graph contrastive learning (GCL) is a newly popular paradigm for self-supervised graph representation learning and offers an alternative to reconstruction-based methods.However, it is not well understood what conditions a task must satisfy such that a given paradigm is better suited. In this paper, we investigate the role of dataset properties and augmentation strategies on the success of GCL and reconstruction-based approaches. Using the recent population augmentation graph-based analysis of self-supervised learning, we show theoretically that the success of GCL with popular augmentations is bounded by the graph edit distance between different classes. Next, we introduce a synthetic data generation process that systematically controls the amount of style vs. content in each sample- i.e. information that is irrelevant vs. relevant to the downstream task- to elucidate how graph representation learning methods perform under different dataset conditions. We empirically show that reconstruction approaches perform better when the style vs. content ratio is low and GCL with popular augmentations benefits from moderate style. Our results provide a general, systematic framework for analyzing different graph representation learning methods and demonstrate when a given approach is expected to perform well.
****
Learning to Shape Rewards using a Game of Two Partners
David Henry Mguni,Jianhong Wang,Taher Jafferjee,Nicolas Perez-Nieves,Wenbin Song,Feifei Tong,Hui Chen,Jiangcheng Zhu,Yaodong Yang,Jun Wang
Reward shaping (RS) is a powerful method in reinforcement learning (RL) for overcoming the problem of sparse or uninformative rewards. However, RS typically relies on manually engineered shaping-reward functions whose construction is time consuming and error-prone. It also requires domain knowledge which runs contrary to the goal of autonomous learning. We introduce Reinforcement Learning Optimal Shaping Algorithm (ROSA), an automated RS framework in which the shaping reward function is constructed in a novel Markov game between two agents. A reward-shaping agent (Shaper) uses switching controls to determine which states to add shaping rewards and their optimal values while the other agent (Controller) learns the optimal policy for the task using these shaped rewards. We prove that ROSA, which easily adopts existing RL algorithms, learns to construct a shaping reward function that is tailored to the task thus ensuring efficient convergence to high performance policies. We demonstrate ROSA’s congenial properties in three carefully designed experiments and show its superior performance against state-of-the-art RS algorithms in challenging sparse reward environments.
****
Leveraging unlabeled data to predict out-of-distribution performance
Saurabh Garg,Sivaraman Balakrishnan,Zachary Chase Lipton,Behnam Neyshabur,Hanie Sedghi
Real-world machine learning deployments are characterized by mismatches between the source (training) and target (test) distributions
that may cause performance drops. In this work, we investigate methods for predicting the target domain accuracy using only labeled source data and unlabeled target data. We propose Average Thresholded Confidence (ATC), a practical method that learns a \emph{threshold} on the model's confidence, predicting accuracy as the fraction of unlabeled examples for which model confidence exceeds that threshold. ATC outperforms previous methods across several model architectures, types of distribution shifts (e.g., due to synthetic corruptions, dataset reproduction, or novel subpopulations), and datasets (\textsc{Wilds}-FMoW, ImageNet, \breeds, CIFAR, and MNIST). In our experiments, ATC estimates target performance $2\text{--}4\times$ more accurately than prior methods. We also explore the theoretical foundations of the problem, proving that, in general, identifying the accuracy is just as hard as identifying the optimal predictor and thus, the efficacy of any method rests upon (perhaps unstated) assumptions on the nature of the shif

t. Finally, analyzing our method on some toy distributions, we provide insights concerning when it works.

****

Understanding and Improving Robustness of Vision Transformers through Patch-based Negative Augmentation

Yao Qin,Chiyuan Zhang,Ting Chen,Balaji Lakshminarayanan,Alex Beutel,Xuezhi Wang

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 across a wide set of ImageNet based robustness benchmarks. Furthermore, we find our patch-based negative augmentation are complementary to traditional (positive) data augmentation, and together boost the performance further.

****

Enforcing physics-based algebraic constraints for inference of PDE models on unstructured grids

Valerii Iakovlev,Markus Heinonen,Harri Lähdesmäki

Data-driven neural network models have recently shown great success in modelling and learning complex PDE systems. Several works have proposed approaches to include specific physics-based constraints to avoid unrealistic modelling outcomes. While previous works focused on specific constraints and uniform spatial grids, we propose a novel approach for enforcing general pointwise, differential and integral constraints on unstructured spatial grids. The method is based on representing a black-box PDE model's output in terms of a function approximation and enforcing constraints directly on that function. We demonstrate applicability of our approach in learning PDE-driven systems and generating spatial fields with GANs, both on free-form spatial and temporal domains, and show how both kinds of models benefit from incorporation of physics-based constraints.

****

VC dimension of partially quantized neural networks in the overparametrized regime

Yutong Wang,Clayton Scott

Vapnik-Chervonenkis (VC) theory has so far been unable to explain the small generalization error of overparametrized neural networks. Indeed, existing applications of VC theory to large networks obtain upper bounds on VC dimension that are proportional to the number of weights, and for a large class of networks, these upper bound are known to be tight. In this work, we focus on a class of partially quantized networks that we refer to as hyperplane arrangement neural networks (HANNs). Using a sample compression analysis, we show that HANNs can have VC dimension significantly smaller than the number of weights, while being highly expressive. In particular, empirical risk minimization over HANNs in the overparametrized regime achieves the minimax rate for classification with Lipschitz posterior class probability. We further demonstrate the expressivity of HANNs empirically. On a panel of 121 UCI datasets, overparametrized HANNs are able to match the performance of state-of-the-art full-precision models.

****

Geometric Algebra Attention Networks for Small Point Clouds

Matthew Spellings

Much of the success of deep learning is drawn from building architectures that properly respect underlying symmetry and structure in the data on which they operate—a set of considerations that have been united under the banner of geometric deep learning. Often problems in the physical sciences deal with relatively small sets of points in two- or three-dimensional space wherein translation, rotation, and permutation equivariance are important or even vital for models to be useful in practice. In this work, we present rotation- and permutation-equivariant architectures for deep learning on these small point clouds, composed of a set of products of terms from the geometric algebra and reductions over those products using an attention mechanism. The geometric algebra provides valuable mathematical structure by which to combine vector, scalar, and other types of geometric inputs in a systematic way to account for rotation invariance or covariance, while attention yields a powerful way to impose permutation equivariance. We demonstrate the usefulness of these architectures by training models to solve sample problems relevant to physics, chemistry, and biology.

****

rQdia: Regularizing Q-Value Distributions With Image Augmentation

Samuel Lerman,Jing Bi,Chenliang Xu

rQdia (pronounced “Arcadia”) regularizes Q-value distributions with augmented images in pixel-based deep reinforcement learning. With a simple auxiliary loss, that equalizes these distributions via MSE, rQdia boosts DrQ and SAC on 9/12 and 10/12 tasks respectively in the MuJoCo Continuous Control Suite from pixels, and Data-Efficient Rainbow on 18/26 Atari Arcade environments. Gains are measured in both sample efficiency and longer-term training. Moreover, the addition of rQdia finally propels model-free continuous control from pixels over the state encoding baseline. Additional results, namely more random seeds, pending.

****

Optimal Representations for Covariate Shift

Yangjun Ruan,Yann Dubois,Chris J. Maddison

Machine learning systems often experience a distribution shift between training and testing. In this paper, we introduce a simple variational objective whose optima are exactly the set of all representations on which risk minimizers are guaranteed to be robust to any distribution shift that preserves the Bayes predictor, e.g., covariate shifts. Our objective has two components. First, a representation must remain discriminative for the task, i.e., some predictor must be able to simultaneously minimize the source and target risk. Second, the representation's marginal support needs to be the same across source and target. We make this practical by designing self-supervised objectives that only use unlabelled data and augmentations to train robust representations.

Our objectives give insights into the robustness of CLIP, and further improve CLIP's representations to achieve SOTA results on DomainBed.

****

Fortuitous Forgetting in Connectionist Networks

Hattie Zhou,Ankit Vani,Hugo Larochelle,Aaron Courville

Forgetting is often seen as an unwanted characteristic in both human and machine learning. However, we propose that forgetting can in fact be favorable to learning. We introduce forget-and-relearn as a powerful paradigm for shaping the learning trajectories of artificial neural networks. In this process, the forgetting step selectively removes undesirable information from the model, and the relearning step reinforces features that are consistently useful under different conditions. The forget-and-relearn framework unifies many existing iterative training algorithms in the image classification and language emergence literature, and allows us to understand the success of these algorithms in terms of the disproportionate forgetting of undesirable information. We leverage this understanding to improve upon existing algorithms by designing more targeted forgetting operations. Insights from our analysis provide a coherent view on the dynamics of iterative training in neural networks and offer a clear path towards performance impro

vements.
****

Understanding Latent Correlation-Based Multiview Learning and Self-Supervision: An Identifiability Perspective

Qi Lyu,Xiao Fu,Weiran Wang,Songtao Lu

Multiple views of data, both naturally acquired (e.g., image and audio) and artificially produced (e.g., via adding different noise to data samples), have proven useful in enhancing representation learning. Natural views are often handled by multiview analysis tools, e.g., (deep) canonical correlation analysis [(D)CCA], while the artificial ones are frequently used in self-supervised learning (SSL) paradigms, e.g., BYOL and Barlow Twins. Both types of approaches often involve learning neural feature extractors such that the embeddings of data exhibit high cross-view correlations. Although intuitive, the effectiveness of correlation-based neural embedding is mostly empirically validated.

This work aims to understand latent correlation maximization-based deep multiview learning from a latent component identification viewpoint. An intuitive generative model of multiview data is adopted, where the views are different nonlinear mixtures of shared and private components. Since the shared components are view/distortion-invariant, representing the data using such components is believed to reveal the identity of the samples effectively and robustly. Under this model, latent correlation maximization is shown to guarantee the extraction of the shared components across views (up to certain ambiguities). In addition, it is further shown that the private information in each view can be provably disentangled from the shared using proper regularization design. A finite sample analysis, which has been rare in nonlinear mixture identifiability study, is also presented. The theoretical results and newly designed regularization are tested on a series of tasks.
****

EigenGame Unloaded: When playing games is better than optimizing

Ian Gemp,Brian McWilliams,Claire Vernade,Thore Graepel

We build on the recently proposed EigenGame that views eigendecomposition as a competitive game. EigenGame's updates are biased if computed using minibatches of data, which hinders convergence and more sophisticated parallelism in the stochastic setting. In this work, we propose an unbiased stochastic update that is asymptotically equivalent to EigenGame, enjoys greater parallelism allowing computation on datasets of larger sample sizes, and outperforms EigenGame in experiments. We present applications to finding the principal components of massive datasets and performing spectral clustering of graphs. We analyze and discuss our proposed update in the context of EigenGame and the shift in perspective from optimization to games.
****

HD-cos Networks: Efficient Neural Architechtures for Secure Multi-Party Computation

Wittawat Jitkrittum,Michal Lukasik,Ananda Theertha Suresh,Felix Yu,Gang Wang

Multi-party computation (MPC) is a branch of cryptography where multiple non-colluding parties execute a well designed protocol to securely compute a function. With the non-colluding party assumption, MPC has a cryptographic guarantee that the parties will not learn sensitive information from the computation process, making it an appealing framework for applications that involve privacy-sensitive user data.

In this paper, we study training and inference of neural networks under the MPC setup. This is challenging because the elementary operations of neural networks such as the ReLU activation function and matrix-vector multiplications are very expensive to compute due to the added multi-party communication overhead.

To address this, we propose the HD-cos network that uses 1) cosine as activation function, 2) the Hadamard-Diagonal transformation to replace the unstructured linear transformations. We show that both of the approaches enjoy strong theoretical motivations and efficient computation under the MPC setup. We demonstrate on multiple public datasets that HD-cos matches the quality of the more expensive baselines.

****

WaveSense: Efficient Temporal Convolutions with Spiking Neural Networks for Keyword Spotting

Philipp Weidel,Sadique Sheik

Ultra-low power local signal processing is a crucial aspect for edge applications on always-on devices.

Neuromorphic processors emulating spiking neural networks show great computational power while fulfilling the limited power budget as needed in this domain.

In this work we propose spiking neural dynamics as a natural alternative to dilated temporal convolutions. We extend this idea to WaveSense, a spiking neural network inspired by the WaveNet architecture.

WaveSense uses simple neural dynamics, fixed time-constants and a simple feed-forward architecture and hence is particularly well suited for a neuromorphic implementation.

We test the capabilities of this model on several datasets for keyword-spotting. The results show that the proposed network beats the state of the art of other spiking neural networks and reaches near state-of-the-art performance of artificial neural networks such as CNNs and LSTMs.

****

GRAPHIX: A Pre-trained Graph Edit Model for Automated Program Repair

Thanh V Nguyen,Srinivasan H. Sengamedu

We present GRAPHIX, a pre-trained graph edit model for automatically detecting and fixing bugs and code quality issues in Java programs. Unlike sequence-to-sequence models, GRAPHIX leverages the abstract syntax structure of code and represents the code using a multi-head graph encoder. Along with an autoregressive tree decoder, the model learns to perform graph edit actions for automated program repair. We devise a novel pre-training strategy for GRAPHIX, namely deleted sub-tree reconstruction, to enrich the model with implicit knowledge of program structures from unlabeled source code. The pre-training objective is made consistent with the bug fixing task to facilitate the downstream learning. We evaluate GRAPHIX on the Patches in The Wild Java benchmark, using both abstract and concrete code. Experimental results show that GRAPHIX significantly outperforms a wide range of baselines including CodeBERT and BART and is as competitive as other state-of-the-art pre-trained Transformer models despite using one order of magnitude fewer parameters. Further analysis demonstrates strong inductive biases of GRAPHIX in learning meaningful structural and semantic code patterns, both in abstract and concrete source code.

****

Hybrid Cloud-Edge Networks for Efficient Inference

Anil Kag,Igor Fedorov,Aditya Gangrade,Paul Whatmough,Venkatesh Saligrama

Although deep neural networks (DNNs) achieve state-of-the-art accuracy on large-scale and fine-grained prediction tasks, they are high capacity models and often cannot be deployed on edge devices. As such, two distinct paradigms have emerged in parallel: 1) edge device inference for low-level tasks, 2) cloud-based inference for large-scale tasks. We propose a novel hybrid option, which marries these extremes and seeks to bring the latency and computational cost benefits of edge device inference to tasks currently deployed in the cloud. Our proposed method is an end-to-end approach, and involves architecting and training two networks in tandem. The first network is a low-capacity network that can be deployed on an edge device, whereas the second is a high-capacity network deployed in the cloud. When the edge device encounters challenging inputs, these inputs are transmitted and processed on the cloud. Empirically, on the ImageNet classification dataset, our proposed method leads to substantial decrease in the number of floating point operations (FLOPs) used compared to a well-designed high-capacity network, while suffering no excess classification loss. A novel aspect of our method is that, by allowing abstentions on a small fraction of examples ($<20\%$), we can increase accuracy without increasing the edge device memory and FLOPs substantially (up to 7\% higher accuracy and 3X fewer FLOPs on ImageNet with 80\% coverage), relative to MobileNetV3 architectures.

****

Contextualized Scene Imagination for Generative Commonsense Reasoning

PeiFeng Wang,Jonathan Zamora,Junfeng Liu,Filip Ilievski,Muhao Chen,Xiang Ren

Humans use natural language to compose common concepts from their environment into plausible, day-to-day scene descriptions. However, such generative commonsense reasoning (GCSR) skills are lacking in state-of-the-art text generation methods. Descriptive sentences about arbitrary concepts generated by neural text generation models (e.g., pre-trained text-to-text Transformers) are often grammatically fluent but may not correspond to human common sense, largely due to their lack of mechanisms to capture concept relations, to identify implicit concepts, and to perform generalizable reasoning about unseen concept compositions. In this paper, we propose an Imagine-and-Verbalize (I\&V) method, which learns to imagine a relational scene knowledge graph (SKG) with relations between the input concepts, and leverage the SKG as a constraint when generating a plausible scene description. We collect and harmonize a set of knowledge resources from different domains and modalities, providing a rich auxiliary supervision signal for I\&V. The experiments demonstrate the effectiveness of I\&V in improving language models on both concept-to-sentence and concept-to-story generation tasks, while enabling the model to learn well from fewer task examples and generate SKGs that make common sense to human annotators.

****

Scene Transformer: A unified architecture for predicting future trajectories of multiple agents

Jiquan Ngiam,Vijay Vasudevan,Benjamin Caine,Zhengdong Zhang,Hao-Tien Lewis Chiang,Jeffrey Ling,Rebecca Roelofs,Alex Bewley,Chenxi Liu,Ashish Venugopal,David J Weiss,Ben Sapp,Zhifeng Chen,Jonathon Shlens

Predicting the motion of multiple agents is necessary for planning in dynamic environments. This task is challenging for autonomous driving since agents (e.g., vehicles and pedestrians) and their associated behaviors may be diverse and influence one another. Most prior work have focused on predicting independent futures for each agent based on all past motion, and planning against these independent predictions. However, planning against independent predictions can make it challenging to represent the future interaction possibilities between different agents, leading to sub-optimal planning. In this work, we formulate a model for predicting the behavior of all agents jointly, producing consistent futures that account for interactions between agents. Inspired by recent language modeling approaches, we use a masking strategy as the query to our model, enabling one to invoke a single model to predict agent behavior in many ways, such as potentially conditioned on the goal or full future trajectory of the autonomous vehicle or the behavior of other agents in the environment. Our model architecture employs attention to combine features across road elements, agent interactions, and time steps. We evaluate our approach on autonomous driving datasets for both marginal and joint motion prediction, and achieve state of the art performance across two popular datasets. Through combining a scene-centric approach, agent permutation equivariant model, and a sequence masking strategy, we show that our model can unify a variety of motion prediction tasks from joint motion predictions to conditioned prediction.

****

DISSECT: Disentangled Simultaneous Explanations via Concept Traversals

Asma Ghandeharioun,Been Kim,Chun-Liang Li,Brendan Jou,Brian Eoff,Rosalind Picard

Explaining deep learning model inferences is a promising venue for scientific understanding, improving safety, uncovering hidden biases, evaluating fairness, and beyond, as argued by many scholars. One of the principal benefits of counterfactual explanations is allowing users to explore "what-if" scenarios through what does not and cannot exist in the data, a quality that many other forms of explanation such as heatmaps and influence functions are inherently incapable of doing. However, most previous work on generative explainability cannot disentangle important concepts effectively, produces unrealistic examples, or fails to retain relevant information. We propose a novel approach, DISSECT, that jointly trains a generator, a discriminator, and a concept disentangler to overcome such chall

enges using little supervision. DISSECT generates Concept Traversals (CTs), defined as a sequence of generated examples with increasing degrees of concepts that influence a classifier's decision. By training a generative model from a classifier's signal, DISSECT offers a way to discover a classifier's inherent "notion" of distinct concepts automatically rather than rely on user-predefined concepts. We show that DISSECT produces CTs that (1) disentangle several concepts, (2) are influential to a classifier's decision and are coupled to its reasoning due to joint training (3), are realistic, (4) preserve relevant information, and (5) are stable across similar inputs. We validate DISSECT on several challenging synthetic and realistic datasets where previous methods fall short of satisfying desirable criteria for interpretability and show that it performs consistently well. Finally, we present experiments showing applications of DISSECT for detecting potential biases of a classifier and identifying spurious artifacts that impact predictions.
****
Heteroscedastic Temporal Variational Autoencoder For Irregularly Sampled Time Series
Satya Narayan Shukla,Benjamin Marlin
Irregularly sampled time series commonly occur in several domains where they present a significant challenge to standard deep learning models. In this paper, we propose a new deep learning framework for probabilistic interpolation of irregularly sampled time series that we call the Heteroscedastic Temporal Variational Autoencoder (HeTVAE). HeTVAE includes a novel input layer to encode information about input observation sparsity, a temporal VAE architecture to propagate uncertainty due to input sparsity, and a heteroscedastic output layer to enable variable uncertainty in the output interpolations. Our results show that the proposed architecture is better able to reflect variable uncertainty through time due to sparse and irregular sampling than a range of baseline and traditional models, as well as recently proposed deep latent variable models that use homoscedastic output layers.
****
A Neural Tangent Kernel Perspective of Infinite Tree Ensembles
Ryuichi Kanoh,Mahito Sugiyama
In practical situations, the tree ensemble is one of the most popular models along with neural networks. A soft tree is a variant of a decision tree. Instead of using a greedy method for searching splitting rules, the soft tree is trained using a gradient method in which the entire splitting operation is formulated in a differentiable form. Although ensembles of such soft trees have been used increasingly in recent years, little theoretical work has been done to understand their behavior. By considering an ensemble of infinite soft trees, this paper introduces and studies the Tree Neural Tangent Kernel (TNTK), which provides new insights into the behavior of the infinite ensemble of soft trees. Using the TNTK, we theoretically identify several non-trivial properties, such as global convergence of the training, the equivalence of the oblivious tree structure, and the degeneracy of the TNTK induced by the deepening of the trees.
****
AlphaZero-based Proof Cost Network to Aid Game Solving
Ti-Rong Wu,Chung-Chin Shih,Ting Han Wei,Meng-Yu Tsai,Wei-Yuan Hsu,I-Chen Wu
The AlphaZero algorithm learns and plays games without hand-crafted expert knowledge. However, since its objective is to play well, we hypothesize that a better objective can be defined for the related but separate task of solving games. This paper proposes a novel approach to solving problems by modifying the training target of the AlphaZero algorithm, such that it prioritizes solving the game quickly, rather than winning. We train a Proof Cost Network (PCN), where proof cost is a heuristic that estimates the amount of work required to solve problems. This matches the general concept of the so-called proof number from proof number search, which has been shown to be well-suited for game solving. We propose two specific training targets. The first finds the shortest path to a solution, while the second estimates the proof cost. We conduct experiments on solving 15x15 Gomoku and 9x9 Killall-Go problems with both MCTS-based and FDFPN solvers. Compar

isons between using AlphaZero networks and PCN as heuristics show that PCN can solve more problems.
****

STRIC: Stacked Residuals of Interpretable Components for Time Series Anomaly Detection

Luca Zancato,Alessandro Achille,Giovanni Paolini,Alessandro Chiuso,Stefano Soatto

We present a residual-style architecture for interpretable forecasting and anomaly detection in multivariate time series.

Our architecture is composed of stacked residual blocks designed to separate components of the signal such as trends, seasonality, and linear dynamics.

These are followed by a Temporal Convolutional Network (TCN) that can freely model the remaining components and can aggregate global statistics from different time series as context for the local predictions of each time series. The architecture can be trained end-to-end and automatically adapts to the time scale of the signals.

After modeling the signals, we use an anomaly detection system based on the classic CUMSUM algorithm and a variational approximation of the f-divergence to detect both isolated point anomalies and change-points in statistics of the signals.

Our method outperforms state-of-the-art robust statistical methods on typical time series benchmarks where deep networks usually underperform. To further illustrate the general applicability of our method, we show that it can be successfully employed on complex data such as text embeddings of newspaper articles.
****

Escaping Saddle Points in Nonconvex Minimax Optimization via Cubic-Regularized Gradient Descent-Ascent

Ziyi Chen,Qunwei Li,Yi Zhou

The gradient descent-ascent (GDA) algorithm has been widely applied to solve nonconvex minimax optimization problems. However, the existing GDA-type algorithms can only find first-order stationary points of the envelope function of nonconvex minimax optimization problems, which does not rule out the possibility to get stuck at suboptimal saddle points. In this paper, we develop Cubic-GDA -- the first GDA-type algorithm for escaping strict saddle points in nonconvex-strongly-concave minimax optimization. Specifically, the algorithm uses gradient ascent to estimate the second-order information of the minimax objective function, and it leverages the cubic regularization technique to efficiently escape the strict saddle points. Under standard smoothness assumptions on the objective function, we show that Cubic-GDA admits an intrinsic potential function whose value monotonically decreases in the minimax optimization process. Such a property leads to a desired global convergence of Cubic-GDA to a second-order stationary point at a sublinear rate. Moreover, we analyze the convergence rate of Cubic-GDA in the full spectrum of a gradient dominant-type nonconvex geometry. Our result shows that Cubic-GDA achieves an orderwise faster convergence rate than the standard GDA for a wide spectrum of gradient dominant geometry. Our study bridges minimax optimization with second-order optimization and may inspire new developments along this direction.
****

Image2Point: 3D Point-Cloud Understanding with 2D Image Pretrained Models

Chenfeng Xu,Shijia Yang,Bohan Zhai,Bichen Wu,Xiangyu Yue,Wei Zhan,Peter Vajda,Kurt Keutzer,Masayoshi Tomizuka

3D point-clouds and 2D images are different visual representations of the physical world. While human vision can understand both representations, computer vision models designed for 2D image and 3D point-cloud understanding are quite different.

Our paper explores the potential for transferring between these two representations by empirically investigating the feasibility of the transfer, the benefits of the transfer, and shedding light on why the transfer works.

We discovered that we can indeed use the same architecture and pretrained weights of a neural net model to understand both images and point-clouds. Specifically

, we can transfer the pretrained image model to a point-cloud model by \textit{inflating} 2D convolutional filters to 3D and then \textbf{f}inetuning the \textbf{i}mage-\textbf{p}retrained models (FIP).

We discover that, surprisingly, models with minimal finetuning efforts --- only on input, output, and optionally batch normalization layers, can achieve competitive performance on 3D point-cloud classification, beating a wide range of point-cloud models that adopt task-specific architectures and use a variety of tricks. When finetuning the whole model, the performance further improves significantly. Meanwhile, we also find that FIP improves data efficiency, achieving up to 10.0 points top-1 accuracy gain on few-shot classification. It also speeds up the training of point-cloud models by up to 11.1x to reach a target accuracy.

****

Bayesian Framework for Gradient Leakage

Mislav Balunovic,Dimitar Iliev Dimitrov,Robin Staab,Martin Vechev

Federated learning is an established method for training machine learning models without sharing training data. However, recent work has shown that it cannot guarantee data privacy as shared gradients can still leak sensitive information. To formalize the problem of gradient leakage, we propose a theoretical framework that enables, for the first time, analysis of the Bayes optimal adversary phrased as an optimization problem. We demonstrate that existing leakage attacks can be seen as approximations of this optimal adversary with different assumptions on the probability distributions of the input data and gradients. Our experiments confirm the effectiveness of the Bayes optimal adversary when it has knowledge of the underlying distribution. Further, our experimental evaluation shows that several existing heuristic defenses are not effective against stronger attacks, especially early in the training process. Thus, our findings indicate that the construction of more effective defenses and their evaluation remains an open problem.

****

Universalizing Weak Supervision

Changho Shin,Winfred Li,Harit Vishwakarma,Nicholas Carl Roberts,Frederic Sala

Weak supervision (WS) frameworks are a popular way to bypass hand-labeling large datasets for training data-hungry models.

These approaches synthesize multiple noisy but cheaply-acquired estimates of labels into a set of high-quality pseudo-labels for downstream training. However, the synthesis technique is specific to a particular kind of label, such as binary labels or sequences, and each new label type requires manually designing a new synthesis algorithm. Instead, we propose a universal technique that enables weak supervision over any label type while still offering desirable properties, including practical flexibility, computational efficiency, and theoretical guarantees. We apply this technique to important problems previously not tackled by WS frameworks including learning to rank, regression, and learning in hyperbolic space. Theoretically, our synthesis approach produces a consistent estimators for learning some challenging but important generalizations of the exponential family model. Experimentally, we validate our framework and show improvement over baselines in diverse settings including real-world learning-to-rank and regression problems along with learning on hyperbolic manifolds.

****

Maximum n-times Coverage for Vaccine Design

Ge Liu,Alexander Dimitrakakis,Brandon Carter,David Gifford

We introduce the maximum n-times coverage problem that selects k overlays to maximize the summed coverage of weighted elements, where each element must be covered at least n times. We also define the min-cost n-times coverage problem where the objective is to select the minimum set of overlays such that the sum of the weights of elements that are covered at least n times is at least τ. Maximum n-times coverage is a generalization of the multi-set multi-cover problem, is NP-complete, and is not submodular. We introduce two new practical solutions for n-times coverage based on integer linear programming and sequential greedy optimization. We show that maximum n-times coverage is a natural way to frame peptide vaccine design, and find that it produces a pan-strain COVID-19 vaccine design that is superior to 29 other published designs in predicted population coverage and the expected number of peptides displayed by each individual's HLA molecules.

****

Sample and Communication-Efficient Decentralized Actor-Critic Algorithms with Finite-Time Analysis

Ziyi Chen,Yi Zhou,Rong-Rong Chen,Shaofeng Zou

Actor-critic (AC) algorithms have been widely adopted in decentralized multi-agent systems to learn the optimal joint control policy. However, existing decentralized AC algorithms either do not preserve the privacy of agents or are not sample and communication-efficient. In this work, we develop two decentralized AC and natural AC (NAC) algorithms that are private, and sample and communication-efficient. In both algorithms, agents share noisy information to preserve privacy and adopt mini-batch updates to improve sample and communication efficiency. Particularly for decentralized NAC, we develop a decentralized Markovian SGD algorithm with an adaptive mini-batch size to efficiently compute the natural policy gradient. Under Markovian sampling and linear function approximation, we prove the proposed decentralized AC and NAC algorithms achieve the state-of-the-art sample complexities $\mathcal{O}(\epsilon^{-2}\ln\epsilon^{-1})$ and $\mathcal{O}(\epsilon^{-3}\ln\epsilon^{-1})$, respectively, and the same small communication complexity $\mathcal{O}(\epsilon^{-1}\ln\epsilon^{-1})$. Numerical experiments demonstrate that the proposed algorithms achieve lower sample and communication complexities than the existing decentralized AC algorithm.

****

KL Guided Domain Adaptation

A. Tuan Nguyen,Toan Tran,Yarin Gal,Philip Torr,Atilim Gunes Baydin

Domain adaptation is an important problem and often needed for real-world applications. In this problem, instead of i.i.d. training and testing datapoints, we assume that the source (training) data and the target (testing) data have different distributions. With that setting, the empirical risk minimization training procedure often does not perform well, since it does not account for the change in the distribution. A common approach in the domain adaptation literature is to learn a representation of the input that has the same (marginal) distribution over the source and the target domain. However, these approaches often require additional networks and/or optimizing an adversarial (minimax) objective, which can be very expensive or unstable in practice. To improve upon these marginal alignment techniques, in this paper, we first derive a generalization bound for the target loss based on the training loss and the reverse Kullback-Leibler (KL) divergence between the source and the target representation distributions. Based on this bound, we derive an algorithm that minimizes the KL term to obtain a better generalization to the target domain. We show that with a probabilistic representation network, the KL term can be estimated efficiently via minibatch samples without any additional network or a minimax objective. This leads to a theoretically sound alignment method which is also very efficient and stable in practice. Experimental results also suggest that our method outperforms other representation-alignment approaches.

****

Contextual Multi-Armed Bandit with Communication Constraints

Francesco Pase,Deniz Gunduz,Michele Zorzi

We consider a remote Contextual Multi-Armed Bandit (CMAB) problem, in which the decision-maker observes the context and the reward, but must communicate the actions to be taken by the agents over a rate-limited communication channel. This can model, for example, a personalized ad placement application, where the content owner observes the individual visitors to its website, and hence has the context information, but must convey the ads that must be shown to each visitor to a separate entity that manages the marketing content. In this Rate-Constrained CMAB (RC-CMAB) problem, the constraint on the communication rate between the decision-maker and the agents imposes a trade-off between the number of bits sent per agent and the acquired average reward. We are particularly interested in the sce nario in which the number of agents and the number of possible actions are large, while the communication budget is limited. Consequently, it can be considered as a policy compression problem, where the distortion metric is induced by the learning objectives. We first consider the fundamental information theoretic limits of this problem by letting the number of agents go to infinity, and study the regret that can be achieved. Then, we propose a practical coding scheme, and provide numerical results for the achieved regret.

****

Benchmarking Sample Selection Strategies for Batch Reinforcement Learning

Yuwei Fu,Di Wu,Benoit Boulet

Training sample section techniques, such as prioritized experience replay (PER), have been recognized as of significant importance for online reinforcement learning algorithms. Efficient sample selection can help further improve the learning efficiency and the final learning performance. However, the impact of sample selection for batch reinforcement learning algorithms, where we aim to learn a near-optimal policy exclusively from the offline logged dataset, has not been well studied. In this work, we investigate the application of non-uniform sampling techniques in batch reinforcement learning. In particular, we compare six variants of PER based on various heuristic priority metrics that focus on different aspects of the offline learning setting. These metrics include temporal-difference error, n-step return, self-imitation learning objective, pseudo-count, uncertainty, and likelihood. Through extensive experiments on the standard batch RL datasets, we find that non-uniform sampling is also effective in batch RL settings. Furthermore, there is no single metric that works in all situations. Our findings also show that it is insufficient to avoid the bootstrapping error in batch reinforcement learning by only changing the sampling scheme.

****

Attentional meta-learners for few-shot polythetic classification

Ben Day,Ramon Viñas Torné,Nikola Simidjievski,Pietro Lio

Polythetic classifications, based on shared patterns of features that need neither be universal nor constant among members of a class, are common in the natural world and greatly outnumber monothetic classifications over a set of features. We show that threshold meta-learners, such as Prototypical Networks, require an embedding dimension that is exponential in the number of features to emulate these functions. In contrast, attentional classifiers, such as Matching Networks, are polythetic by default and able to solve these problems with a linear embedding dimension. However, we find that in the presence of task-irrelevant features, inherent to meta-learning problems, attentional models are susceptible to misclassification. To address this challenge, we propose a self-attention feature-selection mechanism that adaptively dilutes non-discriminative features. We demonstrate the effectiveness of our approach in meta-learning Boolean functions, and synthetic and real-world few-shot learning tasks.

****

From Stars to Subgraphs: Uplifting Any GNN with Local Structure Awareness

Lingxiao Zhao,Wei Jin,Leman Akoglu,Neil Shah

Message Passing Neural Networks (MPNNs) are a common type of Graph Neural Network (GNN), in which each node’s representation is computed recursively by aggregating representations (“messages”) from its immediate neighbors akin to a star-shaped pattern. MPNNs are appealing for being efficient and scalable, however their expressiveness is upper-bounded by the 1st-order Weisfeiler-Lehman isomorphism test (1-WL). In response, prior works propose highly expressive models at the cost of scalability and sometimes generalization performance. Our work stands between these two regimes: we introduce a general framework to uplift any MPNN to be more expressive, with limited scalability overhead and greatly improved practical performance. We achieve this by extending local aggregation in MPNNs from star patterns to general subgraph patterns (e.g., k-egonets): in our framework, each node representation is computed as the encoding of a surrounding induced subgraph rather than encoding of immediate neighbors only (i.e. a star). We choose the subgraph encoder to be a GNN (mainly MPNNs, considering scalability) to design a general framework that serves as a wrapper to uplift any GNN. We call our pro posed method GNN-AK (GNN As Kernel), as the framework resembles a convolutional neural network by replacing the kernel with GNNs. Theoretically, we show that our framework is strictly more powerful than 1&2-WL, and is not less powerful than 3-WL. We also design subgraph sampling strategies which greatly reduce memory footprint and improve speed while maintaining performance. Our method sets new state-of-the-art performance by large margins for several well-known graph ML tasks; specifically, 0.08 MAE on ZINC, 74.79% and 86.887% accuracy on CIFAR10 and PATTERN respectively.

****

NETWORK INSENSITIVITY TO PARAMETER NOISE VIA PARAMETER ATTACK DURING TRAINING

Julian Büchel,Fynn Firouz Faber,Dylan Richard Muir

Neuromorphic neural network processors, in the form of compute-in-memory crossbar arrays of memristors, or in the form of subthreshold analog and mixed-signal ASICs, promise enormous advantages in compute density and energy efficiency for NN-based ML tasks. However, these technologies are prone to computational non-idealities, due to process variation and intrinsic device physics. This degrades the task performance of networks deployed to the processor, by introducing parameter noise into the deployed model. While it is possible to calibrate each device, or train networks individually for each processor, these approaches are expensive and impractical for commercial deployment. Alternative methods are therefore needed to train networks that are inherently robust against parameter variation, as a consequence of network architecture and parameters. We present a new network training algorithm that attacks network parameters during training, and promotes robust performance during inference in the face of random parameter variation. Our approach introduces a loss regularization term that penalizes the susceptibility of a network to weight perturbation. We compare against previous approaches for producing parameter insensitivity such as dropout, weight smoothing and introducing parameter noise during training. We show that our approach produces models that are more robust to random mismatch-induced parameter variation as well as to targeted parameter variation. Our approach finds minima in flatter locations in the weight-loss landscape compared with other approaches, highlighting that the networks found by our technique are less sensitive to parameter perturbation. Our work provides an approach to deploy neural network architectures to inference devices that suffer from computational non-idealities, with minimal loss of performance. This method will enable deployment at scale to novel energy-efficient computational substrates, promoting cheaper and more prevalent edge inference.

****

Deconstructing the Inductive Biases of Hamiltonian Neural Networks

Nate Gruver,Marc Anton Finzi,Samuel Don Stanton,Andrew Gordon Wilson

Physics-inspired neural networks (NNs), such as Hamiltonian or Lagrangian NNs, dramatically outperform other learned dynamics models by leveraging strong inductive biases. These models, however, are challenging to apply to many real world systems, such as those that don’t conserve energy or contain contacts, a common setting for robotics and reinforcement learning. In this paper, we examine the inductive biases that make physics-inspired models successful in practice. We show that, contrary to conventional wisdom, the improved generalization of HNNs is the result of modeling acceleration directly and avoiding artificial complexity from the coordinate system, rather than symplectic structure or energy conservation. We show that by relaxing the inductive biases of these models, we can match or exceed performance on energy-conserving systems while dramatically improving performance on practical, non-conservative systems. We extend this approach to constructing transition models for common Mujoco environments, showing that our model can appropriately balance inductive biases with the flexibility required for model-based control.

****

Offline Meta-Reinforcement Learning with Online Self-Supervision

Vitchyr H. Pong,Ashvin Nair,Laura Smith,Catherine Huang,Sergey Levine

Meta-reinforcement learning (RL) methods can meta-train policies that adapt to new tasks with orders of magnitude less data than standard RL, but meta-training itself is costly and time-consuming. If we can meta-train on offline data, then we can reuse the same static dataset, labeled once with rewards for different tasks, to meta-train policies that adapt to a variety of new tasks at meta-test time. Although this capability would make meta-RL a practical tool for real-world use, offline meta-RL presents additional challenges beyond online meta-RL or standard offline RL settings. Meta-RL learns an exploration strategy that collects data for adapting, and also meta-trains a policy that quickly adapts to data from a new task. Since this policy was meta-trained on a fixed, offline dataset, it might behave unpredictably when adapting to data collected by the learned exploration strategy, which differs systematically from the offline data and thus induces distributional shift. We propose a hybrid offline meta-RL algorithm, which uses offline data with rewards to meta-train an adaptive policy, and then collects additional unsupervised online data, without any reward labels to bridge this distribution shift. By not requiring reward labels for online collection, this data can be much cheaper to collect. We compare our method to prior work on offline meta-RL on simulated robot locomotion and manipulation tasks and find that using additional unsupervised online data collection leads to a dramatic improvement in the adaptive capabilities of the meta-trained policies, matching the performance of fully online meta-RL on a range of challenging domains that require generalization to new tasks.

****

Gradient Importance Learning for Incomplete Observations

Qitong Gao,Dong Wang,Joshua David Amason,Siyang Yuan,Chenyang Tao,Ricardo Henao,Majda Hadziahmetovic,Lawrence Carin,Miroslav Pajic

Though recent works have developed methods that can generate estimates (or imputations) of the missing entries in a dataset to facilitate downstream analysis, most depend on assumptions that may not align with real-world applications and could suffer from poor performance in subsequent tasks such as classification. This is particularly true if the data have large missingness rates or a small sample size. More importantly, the imputation error could be propagated into the prediction step that follows, which may constrain the capabilities of the prediction model. In this work, we introduce the gradient importance learning (GIL) method to train multilayer perceptrons (MLPs) and long short-term memories (LSTMs) to directly perform inference from inputs containing missing values without imputation. Specifically, we employ reinforcement learning (RL) to adjust the gradients used to train these models via back-propagation. This allows the model to exploit the underlying information behind missingness patterns. We test the approach on real-world time-series (i.e., MIMIC-III), tabular data obtained from an eye clinic, and a standard dataset (i.e., MNIST), where our imputation-free predictions outperform the traditional two-step imputation-based predictions using state-of-the-art imputation methods.

****

The Role of Learning Regime, Architecture and Dataset Structure on Systematic Generalization in Simple Neural Networks

Devon Jarvis,Richard Klein,Benjamin Rosman,Andrew M Saxe

Humans often systematically generalize in situations where standard deep neural networks do not. Empirical studies have shown that the learning procedure and network architecture can influence systematicity in deep networks, but the underlying reasons for this influence remain unclear. Here we theoretically study the acquisition of systematic knowledge by simple neural networks. We introduce a minimal space of datasets with systematic and non-systematic features in both the input and output. For shallow and deep linear networks, we derive learning trajectories for all datasets in this space. The solutions reveal that both shallow and deep networks rely on non-systematic inputs to the same extent throughout learning, such that even with early stopping, no networks learn a fully systematic mapping. Turning to the impact of architecture, we show that modularity improves extraction of systematic structure, but only achieves perfect systematicity in the trivial setting where systematic mappings are fully segregated from non-systematic information. Finally, we analyze iterated learning, a procedure in which generations of networks learn from languages generated by earlier learners. Here we find that networks with output modularity successfully converge over generations to a fully systematic `language’ starting from any dataset in our space. Our results contribute to clarifying the role of learning regime, architecture, and dataset structure in promoting systematic generalization, and provide theoretical support for empirical observations that iterated learning can improve systematicity.

****

Deep Probability Estimation

Weicheng Zhu,Matan Leibovich,Sheng Liu,Sreyas Mohan,Aakash Kaku,Boyang Yu,Laure Zanna,Narges Razavian,Carlos Fernandez-Granda

Reliable probability estimation is of crucial importance in many real-world applications where there is inherent uncertainty, such as weather forecasting, medical prognosis, or collision avoidance in autonomous vehicles. Probability-estimation models are trained on observed outcomes (e.g. whether it has rained or not, or whether a patient has died or not), because the ground-truth probabilities of the events of interest are typically unknown. The problem is therefore analogous to binary classification, with the important difference that the objective is to estimate probabilities rather than predicting the specific outcome. The goal of this work is to investigate probability estimation from high-dimensional data using deep neural networks. There exist several methods to improve the probabilities generated by these models but they mostly focus on classification problems where the probabilities are related to model uncertainty. In the case of problems with inherent uncertainty, it is challenging to evaluate performance without access to ground-truth probabilities. To address this, we build a synthetic dataset to study and compare different computable metrics. We evaluate existing methods on the synthetic data as well as on three real-world probability estimation tasks, all of which involve inherent uncertainty: precipitation forecasting from radar images, predicting cancer patient survival from histopathology images, and predicting car crashes from dashcam videos. Finally, we also propose a new method for probability estimation using neural networks, which modifies the training process to promote output probabilities that are consistent with empirical probabilities computed from the data. The method outperforms existing approaches on most metrics on the simulated as well as real-world data.

****

Memorizing Transformers

Yuhuai Wu,Markus Norman Rabe,DeLesley Hutchins,Christian Szegedy

Language models typically need to be trained or finetuned in order to acquire new knowledge, which involves updating their weights.

We instead envision language models that can simply read and memorize new data at inference time, thus acquiring new knowledge immediately. In this work, we extend language models with the ability to memorize the internal representations of past inputs. We demonstrate that an approximate $k$NN lookup into a non-differentiable memory of recent (key, value) pairs improves language modeling across various benchmarks and tasks, including generic webtext (C4), math papers (arXiv), books (PG-19), code (Github), as well as formal theorems (Isabelle). We show that the performance steadily improves when we increase the size of memory up to 262K tokens.

On benchmarks including code and mathematics, we find that the model is capable of making use of newly defined functions and theorems during test time.

****

LPRules: Rule Induction in Knowledge Graphs Using Linear Programming

Sanjeeb Dash,Joao Goncalves

Knowledge graph (KG) completion is a well-studied problem in AI. Rule-based methods and embedding-based methods form two of the solution techniques. Rule-based methods learn first-order logic rules that capture existing facts in an input graph and then use these rules for reasoning about missing facts. A major drawback of such methods is the lack of scalability to large datasets. In this paper, we present a simple linear programming (LP) model to choose rules from a list of candidate rules and assign weights to them. For smaller KGs, we use simple heuristics to create the candidate list. For larger KGs, we start with a small initial candidate list, and then use standard column generation ideas to add more rules in order to improve the LP model objective value. To foster interpretability and generalizability, we limit the complexity of the set of chosen rules via explicit constraints, and tune the complexity hyperparameter for individual datasets. We show that our method can obtain state-of-the-art results for three out of four widely used KG datasets, while taking significantly less computing time than other popular rule learners including some based on neuro-symbolic methods. The improved scalability of our method allows us to tackle large datasets such as YAGO3-10.

****

Less data is more: Selecting informative and diverse subsets with balancing constraints

Srikumar Ramalingam,Daniel Glasner,Kaushal Patel,Raviteja Vemulapalli,Sadeep Jayasumana,Sanjiv Kumar

Deep learning has yielded extraordinary results in vision and natural language processing, but this achievement comes at a cost. Most models require enormous resources during training, both in terms of computation and in human labeling effort. We show that we can identify informative and diverse subsets of data that lead to deep learning models with similar performance as the ones trained with the original dataset. Prior methods have exploited diversity and uncertainty in submodular objective functions for choosing subsets. In addition to these measures, we show that balancing constraints on predicted class labels and decision boundaries are beneficial. We propose a novel formulation of these constraints using matroids, an algebraic structure that generalizes linear independence in vector spaces, and present an efficient greedy algorithm with constant approximation guarantees. We outperform competing baselines on standard classification datasets such as CIFAR-10, CIFAR-100, ImageNet, as well as long-tailed datasets such as CIFAR-100-LT.

****

Bootstrapped Meta-Learning

Sebastian Flennerhag,Yannick Schroecker,Tom Zahavy,Hado van Hasselt,David Silver,Satinder Singh

Meta-learning empowers artificial intelligence to increase its efficiency by learning how to learn. Unlocking this potential involves overcoming a challenging meta-optimisation problem. We propose an algorithm that tackles this problem by letting the meta-learner teach itself. The algorithm first bootstraps a target from the meta-learner, then optimises the meta-learner by minimising the distance to that target under a chosen (pseudo-)metric. Focusing on meta-learning with gradients, we establish conditions that guarantee performance improvements and show that metric can be used to control meta-optimisation. Meanwhile, the bootstrapping mechanism can extend the effective meta-learning horizon without requiring backpropagation through all updates. We achieve a new state-of-the art for model-free agents on the Atari ALE benchmark and demonstrate that it yields both performance and efficiency gains in multi-task meta-learning. Finally, we explore how bootstrapping opens up new possibilities and find that it can meta-learn efficient exploration in an epsilon-greedy Q-learning agent - without backpropagating through the update rule.

****

Autoregressive Latent Video Prediction with High-Fidelity Image Generator

Younggyo Seo,Kimin Lee,Fangchen Liu,Stephen James,Pieter Abbeel

Video prediction is an important yet challenging problem; burdened with the tasks of generating future frames and learning environment dynamics. Recently, autoregressive latent video models have proved to be a powerful video prediction tool, by separating the video prediction into two sub-problems: pre-training an image generator model, followed by learning an autoregressive prediction model in the latent space of the image generator. However, successfully generating high-fidelity and high-resolution videos has yet to be seen. In this work, we investigate how to train an autoregressive latent video prediction model capable of predicting high-fidelity future frames with minimal modification to existing models, and produce high-resolution (256x256) videos. Specifically, we scale up prior mod els by employing a high-fidelity image generator (VQ-GAN) with a causal transformer model, and introduce additional techniques of top-k sampling and data augmentation to further improve video prediction quality. Despite the simplicity, the proposed method achieves competitive performance to state-of-the-art approaches on standard video prediction benchmarks with fewer parameters, and enables high-resolution video prediction on complex and large-scale datasets. Videos are available at the anonymized website https://sites.google.com/view/harp-anonymous

****

Do Users Benefit From Interpretable Vision? A User Study, Baseline, And Dataset

Leon Sixt,Martin Schuessler,Oana-Iuliana Popescu,Philipp Weiß,Tim Landgraf

A variety of methods exist to explain image classification models. However, whether they provide any benefit to users over simply comparing various inputs and the model’s respective predictions remains unclear. We conducted a user study (N=240) to test how such a baseline explanation technique performs against concept-based and counterfactual explanations. To this end, we contribute a synthetic dataset generator capable of biasing individual attributes and quantifying their relevance to the model. In a study, we assess if participants can identify the relevant set of attributes compared to the ground-truth. Our results show that the baseline outperformed concept-based explanations. Counterfactual explanations from an invertible neural network performed similarly as the baseline. Still, they allowed users to identify some attributes more accurately. Our results highlight the importance of measuring how well users can reason about biases of a model, rather than solely relying on technical evaluations or proxy tasks. We open-source our study and dataset so it can serve as a blue-print for future studies.

****

Learning Identity-Preserving Transformations on Data Manifolds

Marissa Catherine Connor,Kion Fallah,Christopher John Rozell

Many machine learning techniques incorporate identity-preserving transformations into their models to generalize their performance to previously unseen data. These transformations are typically selected from a set of functions that are known to maintain the identity of an input when applied (e.g., rotation, translation, flipping, and scaling). However, there are many natural variations that cannot be labeled for supervision or defined through examination of the data. As suggested by the manifold hypothesis, many of these natural variations live on or near a low-dimensional, nonlinear manifold. Several techniques represent manifold variations through a set of learned Lie group operators that define directions of motion on the manifold. However theses approaches are limited because they require transformation labels when training their models and they lack a method for determining which regions of the manifold are appropriate for applying each specific operator. We address these limitations by introducing a learning strategy that does not require transformation labels and developing a method that learns the local regions where each operator is likely to be used while preserving the identity of inputs. Experiments on MNIST and Fashion MNIST highlight our model's ability to learn identity-preserving transformations on multi-class datasets. Additionally, we train on CelebA to showcase our model's ability to learn semantically meaningful transformations on complex datasets in an unsupervised manner.

****

Generalized Fourier Features for Coordinate-Based Learning of Functions on Manifolds

Carlos Esteves,Tianjian Lu,Mohammed Suhail,Yi-fan Chen■,Ameesh Makadia

Recently, positional encoding of input coordinates has been found crucial to enable learning of high-frequency functions with multilayer perceptrons taking low-dimensional coordinate values. In this setting, sinusoids are typically used as a basis for the encoding, which is commonly referred to as "Fourier Features". However, using sinusoids as a basis assumes that the input coordinates lie on Euclidean space. In this work, we generalize positional encoding with Fourier features to non-Euclidean manifolds. We find appropriate bases for positional encoding on manifolds through generalizations of Fourier series. By ensuring the encodings lie on a hypersphere and that the appropriate shifts on the manifold preserve inner-products between encodings, our model approximates convolutions on the manifold, according to the neural tangent kernel (NTK) assumptions. We demonstrate our method on various tasks on different manifolds: 1) learning panoramas on the sphere, 2) learning probability distributions on the rotation manifold, 3) learning neural radiance fields on the product of cube and sphere, and 4) learning light fields represented as the product of spheres.

****

Understanding the Variance Collapse of SVGD in High Dimensions

Jimmy Ba,Murat A Erdogdu,Marzyeh Ghassemi,Shengyang Sun,Taiji Suzuki,Denny Wu,Tianzong Zhang

Stein variational gradient descent (SVGD) is a deterministic inference algorithm that evolves a set of particles to fit a target distribution. Despite its computational efficiency, SVGD often underestimates the variance of the target distribution in high dimensions. In this work we attempt to explain the variance collapse in SVGD. On the qualitative side, we compare the SVGD update with gradient descent on the maximum mean discrepancy (MMD) objective; we observe that the variance collapse phenomenon relates to the bias from deterministic updates present in the "driving force" of SVGD, and empirically verify that removal of such bias leads to more accurate variance estimation. On the quantitative side, we demonstrate that the variance collapse of SVGD can be accurately predicted in the proportional asymptotic limit, i.e., when the number of particles n and dimensions d diverge at the same rate. In particular, for learning high-dimensional isotropic Gaussians, we derive the exact equilibrium variance for both SVGD and MMD-descent under certain near-orthogonality assumption on the converged particles, and confirm that SVGD suffers from the "curse of dimensionality".

****

Multi-batch Reinforcement Learning via Sample Transfer and Imitation Learning

Di Wu,Tianyu Li,David Meger,Michael Jenkin,Xue Liu,Gregory Dudek

Reinforcement learning (RL), especially deep reinforcement learning, has achieved impressive performance on different control tasks. Unfortunately, most online reinforcement learning algorithms require a large number of interactions with the environment to learn a reliable control policy. This assumption of the availability of repeated interactions with the environment does not hold for many real-world applications due to safety concerns, the cost/inconvenience related to interactions, or the lack of an accurate simulator to enable effective sim2real training. As a consequence, there has been a surge in research addressing this issue, including batch reinforcement learning. Batch RL aims to learn a good control policy from a previously collected dataset. Most existing batch RL algorithms are designed for a single batch setting and assume that we have a large number of interaction samples in fixed data sets. These assumptions limit the use of batch RL algorithms in the real world. We use transfer learning to address this data efficiency challenge. This approach is evaluated on multiple continuous control tasks against several robust baselines. Compared with other batch RL algorithms, the methods described here can be used to deal with more general real-world scenarios.

****

Task-Agnostic Graph Neural Explanations

Yaochen Xie,Sumeet Katariya,Xianfeng Tang,Edward W Huang,Nikhil Rao,Karthik Subbian,Shuiwang Ji

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 down-stream tasks. To address these limitations, we propose a Task-Agnostic Graph Neural Explainer (TAGE) trained under self-supervision without knowledge about downstream tasks.   TAGE enables the explanation of GNN embedding models without downstream tasks and allows efficient explanation of multitask models. Our extensi ve experiments show that TAGE can significantly speed up the explanation efficiency while achieving explanation quality as good as or even better than current state-of-the-art GNN explanation approaches.

****

Generalisation in Lifelong Reinforcement Learning through Logical Composition

Geraud Nangue Tasse,Steven James,Benjamin Rosman

We leverage logical composition in reinforcement learning to create a framework that enables an agent to autonomously determine whether a new task can be immediately solved using its existing abilities, or whether a task-specific skill should be learned. In the latter case, the proposed algorithm also enables the agent to learn the new task faster by generating an estimate of the optimal policy. Importantly, we provide two main theoretical results: we bound the performance of the transferred policy on a new task, and we give bounds on the necessary and sufficient number of tasks that need to be learned throughout an agent's lifetime to generalise over a distribution. We verify our approach in a series of experiments, where we perform transfer learning both after learning a set of base tasks, and after learning an arbitrary set of tasks. We also demonstrate that, as a side effect of our transfer learning approach, an agent can produce an interpretable Boolean expression of its understanding of the current task. Finally, we demonstrate our approach in the full lifelong setting where an agent receives tasks from an unknown distribution. Starting from scratch, an agent is able to quickly generalise over the task distribution after learning only a few tasks, which are sub-logarithmic in the size of the task space.

****

Learning-Augmented k-means Clustering

Jon C. Ergun,Zhili Feng,Sandeep Silwal,David Woodruff,Samson Zhou

k-means clustering is a well-studied problem due to its wide applicability. Unfortunately, there exist strong theoretical limits on the performance of any algorithm for the k-means problem on worst-case inputs. To overcome this barrier, we consider a scenario where ``advice'' is provided to help perform clustering. Specifically, we consider the k-means problem augmented with a predictor that, given any point, returns its cluster label in an approximately optimal clustering up to some, possibly adversarial, error. We present an algorithm whose performance improves along with the accuracy of the predictor, even though na\"{i}vely following the accurate predictor can still lead to a high clustering cost. Thus if the predictor is sufficiently accurate, we can retrieve a close to optimal clustering with nearly optimal runtime, breaking known computational barriers for algorithms that do not have access to such advice. We evaluate our algorithms on real datasets and show significant improvements in the quality of clustering.

****

PolyLoss: A Polynomial Expansion Perspective of Classification Loss Functions

Zhaoqi Leng,Mingxing Tan,Chenxi Liu,Ekin Dogus Cubuk,Jay Shi,Shuyang Cheng,Dragomir Anguelov

Cross-entropy loss and focal loss are the most common choices when training deep neural networks for classification problems. Generally speaking, however, a good loss function can take on much more flexible forms, and should be tailored for different tasks and datasets. Motivated by how functions can be approximated via a Taylor expansion, we propose a simple framework, named PolyLoss, to view and design loss functions as a linear combination of polynomial functions. Our PolyLoss allows the importance of different polynomial bases to be easily adjusted depending on the targeting tasks and datasets, while naturally subsuming the aforementioned cross-entropy loss and focal loss as special cases. Extensive experimental results show that the optimal choice within the PolyLoss is indeed dependent on the task and dataset. Simply by introducing one extra hyperparameter and adding one line of code, our Poly-1 formulation outperforms the cross-entropy loss and focal loss on 2D image classification, instance segmentation, object detection, and 3D object detection tasks, sometimes by a large margin.

****

Improving Non-Autoregressive Translation Models Without Distillation

Xiao Shi Huang,Felipe Perez,Maksims Volkovs

Transformer-based autoregressive (AR) machine translation models have achieved significant performance improvements, nearing human-level accuracy on some languages. The AR framework translates one token at a time which can be time consuming, especially for long sequences. To accelerate inference, recent work has been exploring non-autoregressive (NAR) approaches that translate blocks of tokens in parallel. Despite significant progress, leading NAR models still lag behind their AR counterparts, and only become competitive when trained with distillation. In this paper we investigate possible reasons behind this performance gap, namely, the indistinguishability of tokens, and mismatch between training and inference. We then propose the Conditional Masked Language Model with Correction (CMLMC) that addresses these problems. Empirically, we show that CMLMC achieves state-of-the-art NAR performance when trained on raw data without distillation and approaches AR performance on multiple datasets. Full code for this work will be released at the time of publication.

****

Characterizing and Measuring the Similarity of Neural Networks with Persistent Homology

David Pérez Fernández,Asier Gutiérrez-Fandiño,Jordi Armengol-Estapé,Marta Villegas

Characterizing the structural properties of neural networks is crucial yet poorly understood, and there are no well-established similarity measures between networks. In this work, we observe that neural networks can be represented as abstract simplicial complex and analyzed using their topological 'fingerprints' via Persistent Homology (PH). We then describe a PH-based representation proposed for characterizing and measuring similarity of neural networks. We empirically show the effectiveness of this representation as a descriptor of different architectures in several datasets. This approach based on Topological Data Analysis is a step towards better understanding neural networks and a useful similarity measure.

****

Explaining Scaling Laws of Neural Network Generalization

Yasaman Bahri,Ethan Dyer,Jared Kaplan,Jaehoon Lee,Utkarsh Sharma

The test loss of well-trained neural networks often follows precise power-law scaling relations with either the size of the training dataset or the number of parameters in the network. We propose a theory that explains and connects these scaling laws. We identify variance-limited and resolution-limited scaling behavior for both dataset and model size, for a total of four scaling regimes. The variance-limited scaling follows simply from the existence of a well-behaved infinite data or infinite width limit, while the resolution-limited regime can be explained by positing that models are effectively resolving a smooth data manifold. In the large width limit, this can be equivalently obtained from the spectrum of certain kernels, and we present evidence that large width and large dataset resolution-limited scaling exponents are related by a duality. We exhibit all four scaling regimes in the controlled setting of large random feature and pretrained models and test the predictions empirically on a range of standard architectures and datasets. We also observe several empirical relationships between datasets and scaling exponents: super-classing image tasks does not change exponents, while changing input distribution (via changing datasets or adding noise) has a strong effect. We further explore the effect of architecture aspect ratio on scaling exponents.

****

A Theory of Tournament Representations

Arun Rajkumar,Vishnu Veerathu,Abdul Bakey Mir

Real-world tournaments are almost always intransitive. Recent works have noted that parametric models which assume d dimensional node representations can effectively model intransitive tournaments. However, nothing is known about the structure of the class of tournaments that arise out of any fixed d dimensional representations. In this work, we develop a novel theory for understanding parametric tournament representations. Our first contribution is to structurally characterize the class of tournaments that arise out of d dimensional representatio ns. We do this by showing that these tournament classes have forbidden configurations that must necessarily be a union of flip classes, a novel way to partition the set of all tournaments. We further characterize rank 2 tournaments completely by showing that the associated forbidden flip class contains just 2 tournaments. Specifically, we show that the rank 2 tournaments are equivalent to locally transitive tournaments. This insight allows us to show that the minimum feedback arc set problem on this tournament class can be solved using the standard Quicksort procedure. We also exhibit specific forbidden configurations for rank 4 tournaments. For a general rank d tournament class, we show that the flip class associated with a coned-doubly regular tournament of size $\mathcal{O}(\sqrt{d})$ must be a forbidden configuration. To answer a dual question, using a celebrated result of Froster, we show a lower bound of $\Theta(\sqrt{n})$ on the minimum dimension needed to represent all tournaments on n nodes. For any given tournament, we show a novel upper bound on the smallest representation dimension that depends on the least size of the number of unique nodes in any feedback arc set of the flip class associated with a tournament. We show how our results also shed light on the upper bound of sign-rank of matrices.

****

Distributionally Robust Learning for Uncertainty Calibration under Domain Shift

Haoxuan Wang,Anqi Liu,Zhiding Yu,Junchi Yan,Yisong Yue,Anima Anandkumar

We propose a framework for learning calibrated uncertainties under domain shifts. We consider the case where the source (training) distribution differs significantly from the target (test) distribution. We detect such domain shifts through the use of binary domain classifier and integrate it with the task network and train them jointly end-to-end. The binary domain classifier yields a density ratio that reflects the closeness of a target (test) sample to the source (training) distribution. We employ it to adjust the uncertainty of prediction in the task network. This idea of using the density ratio is based on the distributionally robust learning (DRL) framework, which accounts for the domain shift through adversarial risk minimization. We demonstrate that our method generates calibrated uncertainties that benefit many downstream tasks, such as unsupervised domain adaptation (UDA) and semi-supervised learning (SSL). In these tasks, methods like self-training and FixMatch use uncertainties to select confident pseudo-labels for re-training. Our experiments show that the introduction of DRL leads to significant improvements in cross-domain performance. We also demonstrate that the estimated density ratios show agreement with the human selection frequencies, suggesting a match with a proxy of human perceived uncertainties.

****

Persistent Homology Captures the Generalization of Neural Networks Without A Validation Set

Asier Gutiérrez-Fandiño,David Pérez Fernández,Jordi Armengol-Estapé,Marta Villegas

The training of neural networks is usually monitored with a validation (holdout) set to estimate the generalization of the model. This is done instead of measuring intrinsic properties of the model to determine whether it is learning appropriately. In this work, we suggest studying the training of neural networks with Algebraic Topology, specifically Persistent Homology (PH). Using simplicial complex representations of neural networks, we study the PH diagram distance evolution on the neural network learning process with different architectures and several datasets. Results show that the PH diagram distance between consecutive neural network states correlates with the validation accuracy, implying that the generalization error of a neural network could be intrinsically estimated without any holdout set.

****

Training sequence labeling models using prior knowledge

Dani El-Ayyass

Sequence labeling task (part-of-speech tagging, named entity recognition) is one of the most common in NLP. At different times, the following architectures were used to solve it: CRF, BiLSTM, BERT (in chronological order). The combined model BiLSTM / BERT + CRF, where the last one is the topmost layer, however, perform s better than just BiLSTM / BERT.

It is common when there is a small amount of labeled data available for the task. Hence it is difficult to train a model with good generalizing capability, so one has to resort to semi-supervised learning approaches. One of them is called pseudo-labeling, the gist of what is increasing the training samples with unlabeled data, but it cannot be used alongside with the CRF layer, as this layer simulates the probability distribution of the entire sequence, not of individual tokens.

In this paper, we propose an alternative to the CRF layer — the Prior Knowledge Layer (PKL), that allows one to obtain probability distributions of each token and also takes into account prior knowledge concerned the structure of label sequences.
****

Assisted Learning for Organizations with Limited Imbalanced Data

Cheng Chen,Jiaying Zhou,Jie Ding,Yi Zhou,Bhavya Kailkhura

We develop an assisted learning framework for assisting organization-level learners to improve their learning performance with limited and imbalanced data. In particular, learners at the organization level usually have sufficient computation resource, but are subject to stringent collaboration policy and information privacy. Their limited imbalanced data often cause biased inference and sub-optimal decision-making. In our assisted learning framework, an organizational learner purchases assistance service from a service provider and aims to enhance its model performance within a few assistance rounds. We develop effective stochastic training algorithms for assisted deep learning and assisted reinforcement learning. Different from existing distributed algorithms that need to frequently transmit gradients or models, our framework allows the learner to only occasionally share information with the service provider, and still achieve a near-oracle model as if all the data were centralized.
****

TRAKR – A reservoir-based tool for fast and accurate classification of neural time-series patterns

Muhammad Furqan Afzal,Christian D Marton,Erin L. Rich,Kanaka Rajan

Neuroscience has seen a dramatic increase in the types of recording modalities and complexity of neural time-series data collected from them. The brain is a highly recurrent system producing rich, complex dynamics that result in different behaviors. Correctly distinguishing such nonlinear neural time series in real-time, especially those with non-obvious links to behavior, could be useful for a wide variety of applications. These include detecting anomalous clinical events such as seizures in epilepsy, and identifying optimal control spaces for brain machine interfaces. It remains challenging to correctly distinguish nonlinear time-series patterns because of the high intrinsic dimensionality of such data, making accurate inference of state changes (for intervention or control) difficult. Simple distance metrics, which can be computed quickly do not yield accurate classifications. On the other end of the spectrum of classification methods, ensembles of classifiers or deep supervised tools offer higher accuracy but are slow, data-intensive, and computationally expensive. We introduce a reservoir-based tool, state tracker (TRAKR), which offers the high accuracy of ensembles or deep supervised methods while preserving the computational benefits of simple distance metrics. After one-shot training, TRAKR can accurately, and in real time, detect deviations in test patterns. By forcing the weighted dynamics of the reservoir to fit a desired pattern directly, we avoid many rounds of expensive optimization. Then, keeping the output weights frozen, we use the error signal generated by the reservoir in response to a particular test pattern as a classification boundary. We show that, using this approach, TRAKR accurately detects changes in synthetic time series. We then compare our tool to several others, showing that it achieves highest classification performance on a benchmark dataset–sequential MNIST–even when corrupted by noise. Additionally, we apply TRAKR to electrocorticography (ECoG) data from the macaque orbitofrontal cortex (OFC), a higher-order b rain region involved in encoding the value of expected outcomes. We show that TRAKR can classify different behaviorally relevant epochs in the neural time series more accurately and efficiently than conventional approaches. Therefore, TRAKR can be used as a fast and accurate tool to distinguish patterns in complex nonlinear time-series data, such as neural recordings.

****

Generating Scenes with Latent Object Models

Patrick Emami,Pan He,Sanjay Ranka,Anand Rangarajan

We introduce a structured latent variable model that learns the underlying data-generating process for a dataset of scenes. Our goals are to obtain a compositional scene representation and to perform scene generation by modeling statistical relationships between scenes as well as between objects within a scene. To make inference tractable, we take inspiration from visual topic models and introduce an interpretable hierarchy of scene-level and object-level latent variables (i.e., slots). Since generating scenes requires modeling dependencies between objects, we cannot make a bag-of-words assumption to simplify inference. Moreover, assuming that slots are generated with an autoregressive prior requires decomposing scenes sequentially during inference which has known limitations. Our approach is to assume that the assignment of objects to slots during generation is a deterministic function of the scene latent variable. This removes the need for sequential scene decomposition and enables us to propose an inference algorithm that uses orderless scene decomposition to indirectly estimate an ordered slot posterior. Qualitative and quantitative analysis establishes that our approach successfully learns a smoothly traversable scene-level latent space. The hierarchy of scene and slot variables improves the ability of slot-based models to generate samples displaying complex object relations. We also demonstrate that the learned hierarchy of representations can be used for a scene-retrieval application with object-centric re-ranking.

****

Convergent and Efficient Deep Q Learning Algorithm

Zhikang T. Wang,Masahito Ueda

Despite the empirical success of the deep Q network (DQN) reinforcement learning algorithm and its variants, DQN is still not well understood and it does not guarantee convergence. In this work, we show that DQN can indeed diverge and cease to operate in realistic settings. Although there exist gradient-based convergent methods, we show that they actually have inherent problems in learning dynamics which cause them to fail even for simple tasks. To overcome these problems, we propose a convergent DQN algorithm (C-DQN) that is guaranteed to converge and can work with large discount factors (0.9998). It learns robustly in difficult settings and can learn several difficult games in the Atari 2600 benchmark that DQN fails to solve.

****

Advancing Nearest Neighbor Explanation-by-Example with Critical Classification Regions

Eoin M. Kenny,Eoin D. Delaney,Mark T. Keane

There is an increasing body of evidence suggesting that post-hoc explanation-by-example with nearest neighbors is a promising solution for the eXplainable Artificial Intelligence (XAI) problem. However, despite being thoroughly researched for decades, such post-hoc methods have never seriously explored how to enhance these explanations by highlighting specific important "parts" in a classification. Here, we propose the notion of Critical Classification Regions (CCRs) to do this, and several possible methods are experimentally compared to determine the best approach for this explanation strategy. CCRs supplement nearest neighbor examples by highlighting similar important "parts" in the image explanation. Experiments across multiple domains show that CCRs represent key features used by the CNN in both the testing and training data. Finally, a suitably-controlled user study (N=163) on ImageNet, shows CCRs improve people’s assessments of the correctness of a CNN’s predictions for difficult classifications due to ambiguity.

****

Trigger Hunting with a Topological Prior for Trojan Detection

Xiaoling Hu,Xiao Lin,Michael Cogswell,Yi Yao,Susmit Jha,Chao Chen

Despite their success and popularity, deep neural networks (DNNs) are vulnerable when facing backdoor attacks. This impedes their wider adoption, especially in mission critical applications. This paper tackles the problem of Trojan detection, namely, identifying Trojaned models – models trained with poisoned data. One popular approach is reverse engineering, i.e., recovering the triggers on a clean image by manipulating the model’s prediction. One major challenge of reverse engineering approach is the enormous search space of triggers. To this end, we propose innovative priors such as diversity and topological simplicity to not only increase the chances of finding the appropriate triggers but also improve the quality of the found triggers. Moreover, by encouraging a diverse set of trigger candidates, our method can perform effectively in cases with unknown target labels. We demonstrate that these priors can significantly improve the quality of the recovered triggers, resulting in substantially improved Trojan detection accuracy as validated on both synthetic and publicly available TrojAI benchmarks.

****

Who Is the Strongest Enemy? Towards Optimal and Efficient Evasion Attacks in Deep RL

Yanchao Sun,Ruijie Zheng,Yongyuan Liang,Furong Huang

Evaluating the worst-case performance of a reinforcement learning (RL) agent under the strongest/optimal adversarial perturbations on state observations (within some constraints) is crucial for understanding the robustness of RL agents. However, finding the optimal adversary is challenging, in terms of both whether we can find the optimal attack and how efficiently we can find it. Existing works on adversarial RL either use heuristics-based methods that may not find the strongest adversary, or directly train an RL-based adversary by treating the agent as a part of the environment, which can find the optimal adversary but may become intractable in a large state space.

This paper introduces a novel attacking method to find the optimal attacks through collaboration between a designed function named "actor" and an RL-based learner named "director'". The actor crafts state perturbations for a given policy perturbation direction, and the director learns to propose the best policy perturbation directions. Our proposed algorithm, PA-AD, is theoretically optimal and significantly more efficient than prior RL-based works in environments with large state spaces. Empirical results show that our proposed PA-AD universally outperforms state-of-the-art attacking methods in various Atari and MuJoCo environments. By applying PA-AD to adversarial training, we achieve state-of-the-art empirical robustness in multiple tasks under strong adversaries.

****

C-MinHash: Improving Minwise Hashing with Circulant Permutation

Xiaoyun Li,Ping Li

Minwise hashing (MinHash) is an important and practical algorithm for generating random hashes to approximate the Jaccard (resemblance) similarity in massive binary (0/1) data. The basic theory of MinHash requires applying hundreds or even thousands of independent random permutations to each data vector in the dataset, in order to obtain reliable results for (e.g.,) building large-scale learning models or approximate near neighbor search in massive data. In this paper, we propose {\bf Circulant MinHash (C-MinHash)} and provide the surprising theoretical results that using only \textbf{two} independent random permutations in a circulant manner leads to uniformly smaller Jaccard estimation variance than that of the classical MinHash with K independent permutations. Experiments are conducted to show the effectiveness of the proposed method. We also analyze a more convenient C-MinHash variant which reduces two permutations to just one, with extensive numerical results to validate that it achieves essentially the same estimation accuracy as using two permutations with rigorous theory.

****

Analyzing Populations of Neural Networks via Dynamical Model Embedding

Jordan Cotler,Kai Sheng Tai,Felipe Hernandez,Blake Elias,David Sussillo

A core challenge in the interpretation of deep neural networks is identifying co mmonalities between the underlying algorithms implemented by distinct networks trained for the same task. Motivated by this problem, we introduce \textsc{Dynamo}, an algorithm that constructs low-dimensional manifolds where each point corresponds to a neural network model, and two points are nearby if the corresponding neural networks enact similar high-level computational processes. \textsc{Dynamo} takes as input a collection of pre-trained neural networks and outputs a \emph{meta-model} that emulates the dynamics of the hidden states as well as the outputs of any model in the collection. The specific model to be emulated is determined by a \emph{model embedding vector} that the meta-model takes as input; these model embedding vectors constitute a manifold corresponding to the given population of models. We apply \textsc{Dynamo} to both RNNs and CNNs, and find that the resulting model embedding manifolds enable novel applications: clustering of neural networks on the basis of their high-level computational processes in a manner that is less sensitive to reparameterization; model averaging of several neural networks trained on the same task to arrive at a new, operable neural network with similar task performance; and semi-supervised learning via optimization on the model embedding manifold. Using a fixed-point analysis of meta-models trained on populations of RNNs, we gain new insights into how similarities of the topology of RNN dynamics correspond to similarities of their high-level computational processes.

****

Neural Temporal Logic Programming

Karan Samel,Zelin Zhao,Binghong Chen,Shuang Li,Dharmashankar Subramanian,Irfan Essa,Le Song

Events across a timeline are a common data representation, seen in different temporal modalities. Individual atomic events can occur in a certain temporal ordering to compose higher level composite events. Examples of a composite event are a patient's medical symptom or a baseball player hitting a home run, caused distinct temporal orderings of patient vitals and player movements respectively. Such salient composite events are provided as labels in temporal datasets and most works optimize models to predict these composite event labels directly. We focus uncovering the underlying atomic events and their relations that lead to the composite events within a noisy temporal data setting. We propose Neural Temporal Logic Programming (Neural TLP) which first learns implicit temporal relations between atomic events and then lifts logic rules for composite events, given only the composite events labels for supervision. This is done through efficiently searching through the combinatorial space of all temporal logic rules in an end-to-end differentiable manner. We evaluate our method on video and on healthcare data where it outperforms the baseline methods for rule discovery.

****

Mind Your Bits and Errors: Prioritizing the Bits that Matter in Variational Autoencoders

Rui Shu,Stefano Ermon

Good likelihoods do not imply great sample quality. However, the precise manner in which models trained to achieve good likelihoods fail at sample quality remains poorly understood. In this work, we consider the task of image generative modeling with variational autoencoders and posit that the nature of high-dimensional image data distributions poses an intrinsic challenge. In particular, much of the entropy in these natural image distributions is attributable to visually imperceptible information. This signal dominates the training objective, giving models an easy way to achieve competitive likelihoods without successful modeling of the visually perceptible bits. Based on this hypothesis, we decompose the task of generative modeling explicitly into two steps: we first prioritize the modeling of visually perceptible information to achieve good sample quality, and then subsequently model the imperceptible information---the bulk of the likelihood signal---to achieve good likelihoods. Our work highlights the well-known adage that "not all bits are created equal" and demonstrates that this property can and should be exploited in the design of variational autoencoders.

****

Chunked Autoregressive GAN for Conditional Waveform Synthesis

Max Morrison,Rithesh Kumar,Kundan Kumar,Prem Seetharaman,Aaron Courville,Yoshua Bengio

Conditional waveform synthesis models learn a distribution of audio waveforms given conditioning such as text, mel-spectrograms, or MIDI. These systems employ deep generative models that model the waveform via either sequential (autoregressive) or parallel (non-autoregressive) sampling. Generative adversarial networks (GANs) have become a common choice for non-autoregressive waveform synthesis. However, state-of-the-art GAN-based models produce artifacts when performing mel-spectrogram inversion. In this paper, we demonstrate that these artifacts correspond with an inability for the generator to learn accurate pitch and periodicity. We show that simple pitch and periodicity conditioning is insufficient for reducing this error relative to using autoregression. We discuss the inductive bias that autoregression provides for learning the relationship between instantaneous frequency and phase, and show that this inductive bias holds even when autoregressively sampling large chunks of the waveform during each forward pass. Relative to prior state-of-the-art GAN-based models, our proposed model, Chunked Autoregressive GAN (CARGAN) reduces pitch error by 40-60%, reduces training time by 58%, maintains a fast inference speed suitable for real-time or interactive applications, and maintains or improves subjective quality.

****

Federated Inference through Aligning Local Representations and Learning a Consensus Graph

Tengfei Ma,Trong Nghia Hoang,Jie Chen

Machine learning is faced with many data challenges when applied in practice. Among them, a notable barrier is that data are distributed and sharing is unrealistic for volume and privacy reasons. Federated learning is a recent formalism to tackle this challenge, so that data owners can develop a common model jointly but use it separately. In this work, we consider a less addressed scenario where a datum consists of multiple parts, each of which belongs to a separate owner. In this scenario, joint efforts are required not only in learning but also in inference. We study \emph{federated inference}, which allows each data owner to learn its own model that captures local data characteristics and copes with data heterogeneity. On the top is a federation of the local data representations, performing global inference that incorporates all distributed parts collectively. To enhance this local--global framework, we propose aligning the ambiguous data representations caused by arbitrary arrangement of neurons in local neural network models, as well as learning a consensus graph among data owners in the global model to improve performance. We demonstrate effectiveness of the proposed framework on four real-life data sets including power grid systems and traffic networks.

****

COPA: Certifying Robust Policies for Offline Reinforcement Learning against Poisoning Attacks

Fan Wu,Linyi Li,Huan Zhang,Bhavya Kailkhura,Krishnaram Kenthapadi,Ding Zhao,Bo Li

As reinforcement learning (RL) has achieved near human-level performance in a variety of tasks, its robustness has raised great attention. While a vast body of research has explored test-time (evasion) attacks in RL and corresponding defenses, its robustness against training-time (poisoning) attacks remains largely unanswered. In this work, we focus on certifying the robustness of offline RL in the presence of poisoning attacks, where a subset of training trajectories could be arbitrarily manipulated. We propose the first certification framework, COPA, to certify the number of poisoning trajectories that can be tolerated regarding different certification criteria. Given the complex structure of RL, we propose two certification criteria: per-state action stability and cumulative reward bound. To further improve the certification, we propose new partition and aggregation protocols to train robust policies. We further prove that some of the proposed certification methods are theoretically tight and some are NP-Complete problems. We leverage COPA to certify three RL environments trained with different algorithms and conclude: (1) The proposed robust aggregation protocols such as temporal aggr egation can significantly improve the certifications; (2) Our certifications for both per-state action stability and cumulative reward bound are efficient and tight; (3) The certification for different training algorithms and environments are different, implying their intrinsic robustness properties. All experimental results are available at https://copa-leaderboard.github.io.

****

Towards Understanding Label Smoothing

Yi Xu,Yuanhong Xu,Qi Qian,Hao Li,Rong Jin

Label smoothing regularization (LSR) is a prevalent component for training deep neural networks and can improve the generalization of models effectively. Although it achieves empirical success, the theoretical understanding about the power of label smoothing, especially about its influence on optimization, is still limited. In this work, we, for the first time, theoretically analyze the convergence behaviors of stochastic gradient descent with label smoothing in deep learning. Our analysis indicates that an appropriate LSR can speed up the convergence by reducing the variance in gradient, which provides a theoretical interpretation on the effectiveness of LSR. Besides, the analysis implies that LSR may slow down the convergence at the end of optimization. Therefore, a novel algorithm, namely Two-Stage LAbel smoothing (TSLA), is proposed to further improve the convergence. With the extensive analysis and experiments on benchmark data sets, the effectiveness of TSLA is verified both theoretically and empirically.

****

Beyond Prioritized Replay: Sampling States in Model-Based Reinforcement Learning via Simulated Priorities

Yangchen Pan,Jincheng Mei,Amir-massoud Farahmand,Martha White,Hengshuai Yao,Mohsen Rohani,Jun Luo

Prioritized Experience Replay (ER) has been empirically shown to improve sample efficiency across many domains and attracted great attention; however, there is little theoretical understanding of why such prioritized sampling helps and its limitations. In this work, we take a deep look at the prioritized ER. In a supervised learning setting, we show the equivalence between the error-based prioritized sampling method for mean squared error and uniform sampling for cubic power loss. We then provide theoretical insight into why it improves convergence rate upon uniform sampling during early learning. Based on the insight, we further point out two limitations of the prioritized ER method: 1) outdated priorities and 2) insufficient coverage of the sample space. To mitigate the limitations, we propose our model-based stochastic gradient Langevin dynamics sampling method. We show that our method does provide states distributed close to an ideal prioritized sampling distribution estimated by the brute-force method, which does not suffer from the two limitations. We conduct experiments on both discrete and continuous control problems to show our approach's efficacy and examine the practical implication of our method in an autonomous driving application.

****

On the Uncomputability of Partition Functions in Energy-Based Sequence Models

Chu-Cheng Lin,Arya D. McCarthy

In this paper, we argue that energy-based sequence models backed by expressive parametric families can result in uncomputable and inapproximable partition functions. Among other things, this makes model selection--and therefore learning model parameters--not only difficult, but generally _undecidable_. The reason is that there are no good deterministic or randomized estimates of partition functions. Specifically, we exhibit a pathological example where under common assumptions, _no_ useful importance sampling estimates of the partition function can guarantee to have variance bounded below a rational number. As alternatives, we consider sequence model families whose partition functions are computable (if they exist), but at the cost of reduced expressiveness. Our theoretical results suggest that statistical procedures with asymptotic guarantees and sheer (but finite) amounts of compute are not the only things that make sequence modeling work; computability concerns must not be neglected as we consider more expressive model parametrizations.

****

ExT5: Towards Extreme Multi-Task Scaling for Transfer Learning
Vamsi Aribandi,Yi Tay,Tal Schuster,Jinfeng Rao,Huaixiu Steven Zheng,Sanket Vaibhav Mehta,Honglei Zhuang,Vinh Q. Tran,Dara Bahri,Jianmo Ni,Jai Gupta,Kai Hui,Sebastian Ruder,Donald Metzler
Despite the recent success of multi-task learning and transfer learning for natural language processing (NLP), few works have systematically studied the effect of scaling up the number of tasks during pre-training. Towards this goal, this paper introduces ExMix (Extreme Mixture): a massive collection of 107 supervised NLP tasks across diverse domains and task-families. Using ExMix, we study the effect of multi-task pre-training at the largest scale to date, and analyze co-training transfer amongst common families of tasks. Through this analysis, we show that manually curating an ideal set of tasks for multi-task pre-training is not straightforward, and that multi-task scaling can vastly improve models on its own. Finally, we propose ExT5: a model pre-trained using a multi-task objective of self-supervised span denoising and supervised ExMix. Via extensive experiments, we show that ExT5 outperforms strong T5 baselines on SuperGLUE, GEM, Rainbow, Closed-Book QA tasks, and several tasks outside of ExMix. ExT5 also significantly improves sample efficiency while pre-training.
****
DP-REC: Private & Communication-Efficient Federated Learning
Aleksei Triastcyn,Matthias Reisser,Christos Louizos
Privacy and communication efficiency are important challenges in federated training of neural networks, and combining them is still an open problem. In this work, we develop a method that unifies highly compressed communication and differential privacy (DP). We introduce a compression technique based on Relative Entropy Coding (REC) to the federated setting. With a minor modification to REC, we obtain a provably differentially private learning algorithm, DP-REC, and show how to compute its privacy guarantees. Our experiments demonstrate that DP-REC drastically reduces communication costs while providing privacy guarantees comparable to the state-of-the-art.
****
Provably Robust Detection of Out-of-distribution Data (almost) for free
Alexander Meinke,Julian Bitterwolf,Matthias Hein
The application of machine learning in safety-critical systems requires a reliable assessment of uncertainy. 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. In this paper we propose a novel method that combines a certifiable OOD detector with a standard classifier from first principles into an OOD aware classifier. This way we achieve the best of two worlds: certifiably adversarially robust OOD detection, even for OOD samples close to the in-distribution, without loss in either prediction accuracy or detection performance for non-manipulated OOD data. Moreover, due to the particular construction our classifier provably avoids the asymptotic overconfidence problem of standard neural networks.
****
k-Median Clustering via Metric Embedding: Towards Better Initialization with Privacy
Chenglin Fan,Ping Li,Xiaoyun Li
In clustering algorithms, the choice of initial centers is crucial for the quality of the learned clusters. We propose a new initialization scheme for the k-median problem in the general metric space (e.g., discrete space induced by graphs), based on the construction of metric embedding tree structure of the data. From the tree, we can extract good initial centers that can be used subsequently for the local search algorithm. Our method, named the HST initialization, can also be easily extended to the setting of differential privacy (DP) to generate private initial centers. Theoretically, the initial centers from HST initialization can achieve lower error than those from another popular initialization method, k-median++, in the non-DP setting. Moreover, with privacy constraint, we show that the error of applying DP local search followed by our private HST initializ

ation improves previous results, and approaches the known lower bound within a small factor. Empirically, experiments are conducted to demonstrate the effectiveness of our methods.

****

Boundary Graph Neural Networks for 3D Simulations

Andreas Mayr,Sebastian Lehner,Arno Mayrhofer,Christoph Kloss,Sepp Hochreiter,Johannes Brandstetter

The abundance of data has given machine learning considerable momentum in natural sciences and engineering. However, the modeling of simulated physical processes remains difficult. A key problem is the correct handling of geometric boundaries. While triangularized geometric boundaries are very common in engineering applications, they are notoriously difficult to model by machine learning approaches due to their heterogeneity with respect to size and orientation. In this work, we introduce Boundary Graph Neural Networks (BGNNs), which dynamically modify graph structures to address boundary conditions. Boundary graph structures are constructed via modifying edges, augmenting node features, and dynamically inserting virtual nodes. The new BGNNs are tested on complex 3D granular flow processes of hoppers and rotating drums which are standard components of industrial machinery. Using precise simulations that are obtained by an expensive and complex discrete element method, BGNNs are evaluated in terms of computational efficiency as well as prediction accuracy of particle flows and mixing entropies. Even if complex boundaries are present, BGNNs are able to accurately reproduce 3D granular flows within simulation uncertainties over hundreds of thousands of simulation timesteps, and most notably particles completely stay within the geometric objects without using handcrafted conditions or restrictions.

****

Provable Adaptation across Multiway Domains via Representation Learning

Zhili Feng,Shaobo Han,Simon Shaolei Du

This paper studies zero-shot domain adaptation where each domain is indexed on a multi-dimensional array, and we only have data from a small subset of domains. Our goal is to produce predictors that perform well on \emph{unseen} domains. We propose a model which consists of a domain-invariant latent representation layer and a domain-specific linear prediction layer with a low-rank tensor structure. Theoretically, we present explicit sample complexity bounds to characterize the prediction error on unseen domains in terms of the number of domains with training data and the number of data per domain. To our knowledge, this is the first finite-sample guarantee for zero-shot domain adaptation. In addition, we provide experiments on two-way MNIST and four-way fiber sensing datasets to demonstrate the effectiveness of our proposed model.

****

Efficient Token Mixing for Transformers via Adaptive Fourier Neural Operators

John Guibas,Morteza Mardani,Zongyi Li,Andrew Tao,Anima Anandkumar,Bryan Catanzaro

Vision transformers have delivered tremendous success in representation learning. This is primarily due to effective token mixing through self attention. However, this scales quadratically with the number of pixels, which becomes infeasible for high-resolution inputs. To cope with this challenge, we propose Adaptive Fourier Neural Operator (AFNO) as an efficient token mixer that learns to mix in the Fourier domain. AFNO is based on a principled foundation of operator learning which allows us to frame token mixing as a continuous global convolution without any dependence on the input resolution. This principle was previously used to design FNO, which solves global convolution efficiently in the Fourier domain and has shown promise in learning challenging PDEs. To handle challenges in visual representation learning such as discontinuities in images and high resolution inputs, we propose principled architectural modifications to FNO which results in memory and computational efficiency. This includes imposing a block-diagonal structure on the channel mixing weights, adaptively sharing weights across tokens, and sparsifying the frequency modes via soft-thresholding and shrinkage. The resulting model is highly parallel with a quasi-linear complexity and has linear memory in the sequence size. AFNO outperforms self-attention mechanisms for few-s hot segmentation in terms of both efficiency and accuracy. For Cityscapes segmentation with the Segformer-B3 backbone, AFNO can handle a sequence size of 65k and outperforms other efficient self-attention mechanisms.

****

Inductive Biases and Variable Creation in Self-Attention Mechanisms

Benjamin L. Edelman,Surbhi Goel,Sham M. Kakade,Cyril Zhang

Self-attention, an architectural motif designed to model long-range interactions in sequential data, has driven numerous recent breakthroughs in natural language processing and beyond. This work provides a theoretical analysis of the inductive biases of self-attention modules, where our focus is to rigorously establish which functions and long-range dependencies self-attention blocks prefer to represent. We show that bounded-norm Transformer layers create sparse variables: they can represent sparse Lipschitz functions of the input sequence, with sample complexity scaling only logarithmically with the context length. We propose new experimental protocols to support the analysis and guide the practice of training Transformers, built around the rich theory of learning sparse Boolean functions.

****

Revisiting Locality-Sensitive Binary Codes from Random Fourier Features

Xiaoyun Li,Ping Li

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 binary codes via stochastic quantization for efficient information retrieval. In this paper, we revisit binary hashing from RFF, and propose SignRFF, a new and simple strategy to extract RFF-based binary codes. We show the locality-sensitivity of SignRFF, and propose a new measure, called ranking efficiency, to theoretically compare different Locality-Sensitive Hashing (LSH) methods with practical implications. Experiments are conducted to show that the proposed 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, we also validate that the proposed ranking efficiency aligns well with the empirical search performance.

****

Defending Against Backdoor Attacks Using Ensembles of Weak Learners

Charles Jin,Melinda Sun,Martin Rinard

A recent line of work has shown that deep networks are susceptible to backdoor data poisoning attacks. Specifically, by injecting a small amount of malicious data into the training distribution, an adversary gains the ability to control the behavior of the model during inference. We propose an iterative training procedure for removing poisoned data from the training set. Our approach consists of two steps. We first train an ensemble of weak learners to automatically discover distinct subpopulations in the training set. We then leverage a boosting framework to exclude the poisoned data and recover the clean data. Our algorithm is based on a novel bootstrapped measure of generalization, which provably separates the clean from the dirty data under mild assumptions. Empirically, our method successfully defends against a state-of-the-art dirty label backdoor attack. We find that our approach significantly outperforms previous defenses.

****

Sample Selection with Uncertainty of Losses for Learning with Noisy Labels

Xiaobo Xia,Tongliang Liu,Bo Han,Mingming Gong,Jun Yu,Gang Niu,Masashi Sugiyama

In learning with noisy labels, the sample selection approach is very popular, which regards small-loss data as correctly labeled data during training. However, losses are generated on-the-fly based on the model being trained with noisy labels, and thus large-loss data are likely but not certain to be incorrect. There are actually two possibilities of a large-loss data point: (a) it is mislabeled, and then its loss decreases slower than other data, since deep neural networks learn patterns first; (b) it belongs to an underrepresented group of data and has not been selected yet. In this paper, we incorporate the uncertainty of losses by adopting interval estimation instead of point estimation of losses, where lower bounds of the confidence intervals of losses derived from distribution-free conc entration inequalities, but not losses themselves, are used for sample selection. In this way, we also give large-loss but less selected data a try; then, we can better distinguish between the cases (a) and (b) by seeing if the losses effectively decrease with the uncertainty after the try. As a result, we can better explore underrepresented data that are correctly labeled but seem to be mislabeled at first glance. Experiments demonstrate that the proposed method is superior to baselines and robust to a broad range of label noise types.

****

Unsupervised Neural Machine Translation with Generative Language Models Only

Jesse Michael Han,Igor Babuschkin,Harrison Edwards,Arvind Neelakantan,Tao Xu,Stanislas Polu,Alex Ray,Pranav Shyam,Aditya Ramesh,Alec Radford,Ilya Sutskever

We show how to derive state-of-the-art unsupervised neural machine translation systems from generatively pre-trained language models. Our method consists of three steps: \emph{few-shot amplification}, \emph{distillation}, and \emph{backtranslation}. We first use the zero-shot translation ability of large pretrained language models to generate translations for a small set of unlabeled sentences. We then amplify these zero-shot translations by using them as few-shot demonstrations for sampling a larger synthetic dataset. This dataset is then distilled by discarding the few-shot demonstrations and then fine-tuning. During backtranslation, we repeatedly generate translations for a set of inputs and then fine-tune a single language model on both directions of the translation task at once, ensuring cycle-consistency by swapping the roles of gold monotext and generated translations when fine-tuning. By using our method to leverage GPT-3's zero-shot translation capability, we achieve a new state-of-the-art in unsupervised translation on the WMT14 English-French benchmark, attaining a BLEU score of 42.1.

****

Perceiver IO: A General Architecture for Structured Inputs & Outputs

Andrew Jaegle,Sebastian Borgeaud,Jean-Baptiste Alayrac,Carl Doersch,Catalin Ionescu,David Ding,Skanda Koppula,Daniel Zoran,Andrew Brock,Evan Shelhamer,Olivier J Henaff,Matthew Botvinick,Andrew Zisserman,Oriol Vinyals,Joao Carreira

A central goal of machine learning is the development of systems that can solve many problems in as many data domains as possible. Current architectures, however, cannot be applied beyond a small set of stereotyped settings, as they bake in domain & task assumptions or scale poorly to large inputs or outputs. In this work, we propose Perceiver IO, a general-purpose architecture that handles data from arbitrary settings while scaling linearly with the size of inputs and outputs. Our model augments the Perceiver with a flexible querying mechanism that enables outputs of various sizes and semantics, doing away with the need for task-specific architecture engineering. The same architecture achieves strong results on tasks spanning natural language and visual understanding, multi-task and multi-modal reasoning, and StarCraft II. As highlights, Perceiver IO outperforms a Transformer-based BERT baseline on the GLUE language benchmark despite removing input tokenization and achieves state-of-the-art performance on Sintel optical flow estimation with no explicit mechanisms for multiscale correspondence.

****

Symmetry-driven graph neural networks

Francesco Farina,Emma Slade

Exploiting symmetries and invariance in data is a powerful, yet not fully exploited, way to achieve better generalisation with more
efficiency. In this paper, we introduce two graph network architectures that are equivariant to several types of transformations affecting the node coordinates. First, we build equivariance to any transformation in the coordinate embeddings that preserves the distance between neighbouring nodes, allowing for equivariance to the Euclidean group. Then, we introduce angle attributes to build equivariance to any angle preserving transformation - thus, to the conformal group. Thanks to their equivariance properties, the proposed models can be vastly more data efficient with respect to classical graph architectures, intrinsically equipped with a better inductive bias and better at generalising. We demonstrate these capabilities on a synthetic dataset composed of n-dimensional geometric objects. Additionally, we provide examples of their limitations when (the right) s

ymmetries are not present in the data.

****

RainNet: A Large-Scale Imagery Dataset for Spatial Precipitation Downscaling

Xuanhong Chen,Kairui Feng,Naiyuan Liu,Yifan Lu,Bingbing Ni,Ziang Liu,Maofeng Liu

Contemporary deep learning frameworks have been applied to solve meteorological problems (\emph{e.g.}, front detection, synthetic radar generation, precipitation nowcasting, \emph{e.t.c.}) and have achieved highly promising results. Spatial precipitation downscaling is one of the most important meteorological problems. 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 \emph{RainNet}, 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 models in precipitation downscaling. Specifically, the precipitation maps carefully collected in RainNet cover various meteorological phenomena (\emph{e.g.}, hurricane, squall, \emph{e.t.c}.), which is of great help to improve the model generalization ability. In addition, the map pairs in RainNet are organized in the form of image sequences (720 maps per month or 1 map/hour), showing complex physical properties, \emph{e.g.}, temporal misalignment, temporal sparse, and fluid properties. Two machine-learning-oriented metrics are specifically introduced to evaluate or verify the comprehensive performance of the trained model, (\emph{e.g.}, prediction maps reconstruction accuracy). To illustrate the applications of RainNet, 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 framework to learn the above characteristics. Extensive experiments demonstrate that the value of RainNet in training and evaluating downscaling models.

****

Data-Driven Offline Optimization for Architecting Hardware Accelerators

Aviral Kumar,Amir Yazdanbakhsh,Milad Hashemi,Kevin Swersky,Sergey Levine

To attain higher efficiency, the industry has gradually reformed towards application-specific hardware accelerators. While such a paradigm shift is already starting to show promising results, designers need to spend considerable manual effort and perform large number of time-consuming simulations to find accelerators that can accelerate multiple target applications while obeying design constraints. Moreover, such a simulation-driven approach must be re-run from scratch every time the set of target applications or design constraints change. An alternative paradigm is to use a data-driven, offline approach that utilizes logged simulation data, to architect hardware accelerators, without needing any form of simulations. Such an approach not only alleviates the need to run time-consuming simulation, but also enables data reuse and applies even when set of target applications changes. In this paper, we develop such a data-driven offline optimization method for designing hardware accelerators, dubbed PRIME, that enjoys all of these properties. Our approach learns a conservative, robust estimate of the desired cost function, utilizes infeasible points and optimizes the design against this estimate without any additional simulator queries during optimization. PRIME architects accelerators---tailored towards both single- and multi-applications---improving performance upon stat-of-the-art simulation-driven methods by about 1.54x and 1.20x, while considerably reducing the required total simulation time by 93% and 99%, respectively. In addition, PRIME also architects effective accelerators for unseen applications in a zero-shot setting, outperforming simulation-based methods by 1.26x.

****

Mixed-Memory RNNs for Learning Long-term Dependencies in Irregularly Sampled Time Series

Mathias Lechner,Ramin Hasani

Recurrent neural networks (RNNs) with continuous-time hidden states are a natural fit for modeling irregularly sampled time series. These models, however, face difficulties when the input data possess long-term dependencies. We prove that s imilar to standard RNNs, the underlying reason for this issue is the vanishing or exploding of the gradient during training. This phenomenon is expressed by the ordinary differential equation (ODE) representation of the hidden state, regardless of the ODE solver's choice. We provide a solution by equipping arbitrary continuous-time networks with a memory compartment separated from its time-continuous state. This way, we encode a continuous-time dynamical flow within the RNN, allowing it to respond to inputs arriving at arbitrary time-lags while ensuring a constant error propagation through the memory path. We call these models Mixed-Memory-RNNs (mmRNNs). We experimentally show that Mixed-Memory-RNNs outperform recently proposed RNN-based counterparts on non-uniformly sampled data with long-term dependencies.

****

Coordination Among Neural Modules Through a Shared Global Workspace

Anirudh Goyal,Aniket Rajiv Didolkar,Alex Lamb,Kartikeya Badola,Nan Rosemary Ke,Nasim Rahaman,Jonathan Binas,Charles Blundell,Michael Curtis Mozer,Yoshua Bengio

Deep learning has seen a movement away from representing examples with a monolithic hidden state towards a richly structured state. For example, Transformers segment by position, and object-centric architectures decompose images into entities. In all these architectures, interactions between different elements are modeled via pairwise interactions: Transformers make use of self-attention to incorporate information from other positions and object-centric architectures make use of graph neural networks to model interactions among entities. We consider how to improve on pairwise interactions in terms of global coordination and a coherent, integrated representation that can be used for downstream tasks. In cognitive science, a global workspace architecture has been proposed in which functionally specialized components share information through a common, bandwidth-limited communication channel. We explore the use of such a communication channel in the context of deep learning for modeling the structure of complex environments. The proposed method includes a shared workspace through which communication among different specialist modules takes place but due to limits on the communication bandwidth, specialist modules must compete for access. We show that capacity limitations have a rational basis in that (1) they encourage specialization and compositionality and (2) they facilitate the synchronization of otherwise independent specialists.

****

Search Spaces for Neural Model Training

Darko Stosic,Dusan Stosic

While larger neural models are pushing the boundaries of what deep learning can do, often more weights are needed to train models rather than to run inference for tasks. This paper seeks to understand this behavior using search spaces -- adding weights creates extra degrees of freedom that form new paths for optimization (or wider search spaces) rendering neural model training more effective. We then show how we can augment search spaces to train sparse models attaining competitive scores across dozens of deep learning workloads. They are also are tolerant of structures targeting current hardware, opening avenues for training and inference acceleration. Our work encourages research to explore beyond massive neural models being used today.

****

Multilevel physics informed neural networks (MPINNs)

Elisa Riccietti,Valentin Mercier,Serge Gratton,Pierre Boudier

In this paper we introduce multilevel physics informed neural networks (MPINNs). Inspired by classical multigrid methods for the solution of linear systems arising from the discretization of PDEs, our MPINNs are based on the classical correction scheme, which represents the solution as the sum of a fine and a coarse term that are optimized in an alternate way. We show that the proposed approach allows to reproduce in the neural network training the classical acceleration effect observed for classical multigrid methods, thus providing a PINN that shows improved performance compared to the state-of-the-art. Thanks to the support of the coarse model, MPINNs provide indeed a faster and improved decrease of the appr oximation error in the case both of elliptic and nonlinear equations.

****

AestheticNet: Reducing bias in facial data sets under ethical considerations

Michael Danner,Muhammad Awais Tanvir Rana,Thomas Weber,Tobias Gerlach,Patrik Huber,Matthias Rätsch,Josef Kittler

Facial Beauty Prediction (FBP) aims to develop a machine that can automatically evaluate facial attractiveness. Usually, these results were highly correlated with human ratings, and therefore also reflected human bias in annotations. Everyone will have biases that are usually subconscious and not easy to notice. Unconscious bias deserves more attention than explicit discrimination. It affects moral judgement and can evade moral responsibility, and we cannot eliminate it completely. A new challenge for scientists is to provide training data and AI algorithms that can withstand distorted information. Our experiments prove that human aesthetic judgements are usually biased. In this work, we introduce AestheticNet, the most advanced attractiveness prediction network, with a Pearson correlation coefficient of 0.9601, which is significantly better than the competition. This network is then used to enrich the training data with synthetic images in order to overwrite the ground truth values with fair assessments.

We propose a new method to generate an unbiased CNN to improve the fairness of machine learning. Prediction and recommender systems based on Artificial Intelligence (AI) technology are widely used in various sectors of industry, such as intelligent recruitment, security, etc. Therefore, their fairness is very important. Our research provides a practical example of how to build a fair and trustable AI.

****

Hierarchically Regularized Deep Forecasting

Biswajit Paria,Rajat Sen,Amr Ahmed,Abhimanyu Das

Hierarchical forecasting is a key problem in many practical multivariate forecasting applications - the goal is to simultaneously predict a large number of correlated time series that are arranged in a pre-specified aggregation hierarchy. The main challenge is to exploit the hierarchical correlations to simultaneously obtain good prediction accuracy for time series at different levels of the hierarchy. In this paper, we propose a new approach for hierarchical forecasting which consists of two components. First, decomposing the time series along a global set of basis time series and modeling hierarchical constraints using the coefficients of the basis decomposition. And second, using a linear autoregressive model with coefficients that vary with time. Unlike past methods, our approach is scalable (inference for a specific time series only needs access to its own history) while also modeling the hierarchical structure via (approximate) coherence constraints among the time series forecasts. We experiment on several public datasets and demonstrate significantly improved overall performance on forecasts at different levels of the hierarchy, compared to existing state-of-the-art hierarchical models.

****

Multi-Agent MDP Homomorphic Networks

Elise van der Pol,Herke van Hoof,Frans A Oliehoek,Max Welling

This paper introduces Multi-Agent MDP Homomorphic Networks, a class of networks that allows distributed execution using only local information, yet is able to share experience between global symmetries in the joint state-action space of cooperative multi-agent systems. In cooperative multi-agent systems, complex symmetries arise between different configurations of the agents and their local observations. For example, consider a group of agents navigating: rotating the state globally results in a permutation of the optimal joint policy. Existing work on symmetries in single agent reinforcement learning can only be generalized to the fully centralized setting, because such approaches rely on the global symmetry in the full state-action spaces, and these can result in correspondences across agents. To encode such symmetries while still allowing distributed execution we propose a factorization that decomposes global symmetries into local transformations. Our proposed factorization allows for distributing the computation that enf orces global symmetries over local agents and local interactions. We introduce a multi-agent equivariant policy network based on this factorization. We show empirically on symmetric multi-agent problems that globally symmetric distributable policies improve data efficiency compared to non-equivariant baselines.

****

Scaling-up Diverse Orthogonal Convolutional Networks by a Paraunitary Framework
Jiahao Su,Wonmin Byeon,Furong Huang

Enforcing orthogonality in neural networks is an antidote for gradient vanishing/exploding problems, sensitivity to adversarial perturbation, and bounding generalization errors. However, many previous approaches are heuristic, and the orthogonality of convolutional layers is not systematically studied. Some of these designs are not exactly orthogonal, while others only consider standard convolutional layers and propose specific classes of their realizations. We propose a theoretical framework for orthogonal convolutional layers to address this problem, establishing the equivalence between diverse orthogonal convolutional layers in the spatial domain and the paraunitary systems in the spectral domain. Since a complete factorization exists for paraunitary systems, any orthogonal convolution layer can be parameterized as convolutions of spatial filters. Our framework endows high expressive power to various convolutional layers while maintaining their exact orthogonality. Furthermore, our layers are memory and computationally efficient for deep networks compared to previous designs. Our versatile framework, for the first time, enables the study of architecture designs for deep orthogonal networks, such as choices of skip connection, initialization, stride, and dilation. Consequently, we scale up orthogonal networks to deep architectures, including ResNet and ShuffleNet, substantially increasing the performance over their shallower counterparts. Finally, we show how to construct residual flows, a flow-based generative model that requires strict Lipschitzness, using our orthogonal networks.

****

Learning-Augmented Sketches for Hessians
Yi Li,Honghao Lin,David Woodruff

Sketching is a dimensionality reduction technique where one compresses a matrix by linear combinations that are typically chosen at random. A line of work has shown how to sketch the Hessian to speed up each iteration in a second order method, but such sketches usually depend only on the matrix at hand, and in a number of cases are even oblivious to the input matrix. One could instead hope to learn a distribution on sketching matrices that is optimized for the specific distribution of input matrices. We show how to design learned sketches for the Hessian in the context of second order methods. We prove that a smaller sketching dimension of the column space of a tall matrix is possible, assuming the knowledge of the indices of the rows of large leverage scores. This would lead to faster convergence of the iterative Hessian sketch procedure. We also design a new objective to learn the sketch, whereby we optimize the subspace embedding property of the sketch. We show empirically that learned sketches, compared with their "non-learned" counterparts, do improve the approximation accuracy for important problems, including LASSO and matrix estimation with nuclear norm constraints.

****

Learning shared neural manifolds from multi-subject FMRI data
Jessie Huang,Erica Lindsey Busch,Tom Wallenstein,Michal Gerasimiuk,Guillaume Lajoie,Guy Wolf,Nicholas Turk-Browne,Smita Krishnaswamy

Functional magnetic resonance imaging (fMRI) is a notoriously noisy measurement of brain activity because of the large variations between individuals, signals marred by environmental differences during collection, and spatiotemporal averaging required by the measurement resolution. In addition, the data is extremely high dimensional, with the space of the activity typically having much lower intrinsic dimension. In order to understand the connection between stimuli of interest and brain activity, and analyze differences and commonalities between subjects, it becomes important to learn a meaningful embedding of the data that denoises, and reveals its intrinsic structure. Specifically, we assume that while noise varies significantly between individuals, true responses to stimuli will share common, low-dimensional features between subjects which are jointly discoverable. Similar approaches have been exploited previously but they have mainly used linear methods such as PCA and shared response modeling (SRM). In contrast, we propose a neural network called MRMD-AE (manifold-regularized multiple-decoder, autoencoder), that learns a common embedding from multiple subjects in an experiment while retaining the ability to decode to individual raw fMRI signals. We show that our learned common space represents an extensible manifold (where new points not seen during training can be mapped), improves the classification accuracy of stimulus features of unseen timepoints, as well as improves cross-subject translation of fMRI signals. We believe this framework can be used for many downstream applications such as guided BCI training in the future.

****

Geometry-Consistent Neural Shape Representation with Implicit Displacement Fields

Wang Yifan,Lukas Rahmann,Olga Sorkine-hornung

We present implicit displacement fields, a novel representation for detailed 3D geometry. Inspired by a classic surface deformation technique, displacement mapping, our method represents a complex surface as a smooth base surface plus a displacement along the base's normal directions, resulting in a frequency-based shape decomposition, where the high-frequency signal is constrained geometrically by the low-frequency signal. Importantly, this disentanglement is unsupervised thanks to a tailored architectural design that has an innate frequency hierarchy by construction. We explore implicit displacement field surface reconstruction and detail transfer
and demonstrate superior representational power, training stability, and generalizability.

****

DR3: Value-Based Deep Reinforcement Learning Requires Explicit Regularization

Aviral Kumar,Rishabh Agarwal,Tengyu Ma,Aaron Courville,George Tucker,Sergey Levine

Despite overparameterization, deep networks trained via supervised learning are surprisingly easy to optimize and exhibit excellent generalization. One hypothesis to explain this is that overparameterized deep networks enjoy the benefits of implicit regularization induced by stochastic gradient descent, which favors parsimonious solutions that generalize well on test inputs. It is reasonable to surmise that deep reinforcement learning (RL) methods could also benefit from this effect. In this paper, we discuss how the implicit regularization effect of SGD seen in supervised learning could in fact be harmful in the offline deep RL setting, leading to poor generalization and degenerate feature representations. Our theoretical analysis shows that when existing models of implicit regularization are applied to temporal difference learning, the resulting derived regularizer favors degenerate solutions with excessive aliasing, in stark contrast to the supervised learning case. We back up these findings empirically, showing that feature representations learned by a deep network value function trained via bootstrapping can indeed become degenerate, aliasing the representations for state-action pairs that appear on either side of the Bellman backup. To address this issue, we derive the form of this implicit regularizer and, inspired by this derivation, propose a simple and effective explicit regularizer, called DR3, that counteracts the undesirable effects of this implicit regularizer. When combined with existing offline RL methods, DR3 substantially improves performance and stability, alleviating unlearning in Atari 2600 games, D4RL domains and robotic manipulation from images.

****

Modeling Label Space Interactions in Multi-label Classification using Box Embeddings

Dhruvesh Patel,Pavitra Dangati,Jay-Yoon Lee,Michael Boratko,Andrew McCallum

Multi-label classification is a challenging structured prediction task in which a set of output class labels are predicted for each input. Real-world datasets often have natural or latent taxonomic relationships between labels, making it desirable for models to employ label representations capable of capturing such tax

onomies. Most existing multi-label classification methods do not do so, resulting in label predictions that are inconsistent with the taxonomic constraints, thus failing to accurately represent the fundamentals of problem setting. In this work, we introduce the multi-label box model (MBM), a multi-label classification method that combines the encoding power of neural networks with the inductive bias and probabilistic semantics of box embeddings (Vilnis, et al 2018). Box embeddings can be understood as trainable Venn-diagrams based on hyper-rectangles. Representing labels by boxes rather than vectors, MBM is able to capture taxonomic relations among labels. Furthermore, since box embeddings allow these relations to be learned by stochastic gradient descent from data, and to be read as calibrated conditional probabilities, our model is endowed with a high degree of interpretability. This interpretability also facilitates the injection of partial information about label-label relationships into model training, to further improve its consistency. We provide theoretical grounding for our method and show experimentally the model's ability to learn the true latent taxonomic structure from data. Through extensive empirical evaluations on both small and large-scale multi-label classification datasets, we show that BBM can significantly improve taxonomic consistency while preserving or surpassing the state-of-the-art predictive performance.
****

Fair Node Representation Learning via Adaptive Data Augmentation
Oyku Deniz Kose,Yanning Shen

Node representation learning has demonstrated its efficacy for various applications on graphs, which leads to increasing attention towards the area. However, fairness is a largely under-explored territory within the field, which may lead to biased results towards underrepresented groups in ensuing tasks. To this end, this work theoretically explains the sources of bias in node representations obtained via Graph Neural Networks (GNNs). Our analysis reveals that both nodal features and graph structure lead to bias in the obtained representations. Building upon the analysis, fairness-aware data augmentation frameworks on nodal features and graph structure are developed to reduce the intrinsic bias. Our analysis and proposed schemes can be readily employed to enhance the fairness of various GNN-based learning mechanisms. Extensive experiments on node classification and link prediction are carried out over real networks in the context of graph contrastive learning. Comparison with multiple benchmarks demonstrates that the proposed augmentation strategies can improve fairness in terms of statistical parity and equal opportunity, while providing comparable utility to state-of-the-art contrastive methods.
****

It Takes Two to Tango: Mixup for Deep Metric Learning
Shashanka Venkataramanan,Bill Psomas,Ewa Kijak,laurent amsaleg,Konstantinos Karantzalos,Yannis Avrithis

Metric learning involves learning a discriminative representation such that embeddings of similar classes are encouraged to be close, while embeddings of dissimilar classes are pushed far apart. State-of-the-art methods focus mostly on sophisticated loss functions or mining strategies. On the one hand, metric learning losses consider two or more examples at a time. On the other hand, modern data augmentation methods for classification consider two or more examples at a time. The combination of the two ideas is under-studied.

In this work, we aim to bridge this gap and improve representations using mixup, which is a powerful data augmentation approach interpolating two or more examples and corresponding target labels at a time. This task is challenging because, unlike classification, the loss functions used in metric learning are not additive over examples, so the idea of interpolating target labels is not straightforward. To the best of our knowledge, we are the first to investigate mixing both examples and target labels for deep metric learning. We develop a generalized formulation that encompasses existing metric learning loss functions and modify it to accommodate for mixup, introducing Metric Mix, or Metrix. We also introduce a new metric---utilization---to demonstrate that by mixing examples during traini ng, we are exploring areas of the embedding space beyond the training classes, thereby improving representations. To validate the effect of improved representations, we show that mixing inputs, intermediate representations or embeddings along with target labels significantly outperforms state-of-the-art metric learning methods on four benchmark deep metric learning datasets.

****

Data Efficient Language-Supervised Zero-Shot Recognition with Optimal Transport Distillation

Bichen Wu,Ruizhe Cheng,Peizhao Zhang,Tianren Gao,Joseph E. Gonzalez,Peter Vajda

Traditional computer vision models are trained to predict a fixed set of predefined categories. Recently, natural language has been shown to be a broader and richer source of supervision that provides finer descriptions to visual concepts than supervised "gold" labels. Previous works, such as CLIP, use InfoNCE loss to train a model to predict the pairing between images and text captions. CLIP, however, is data hungry and requires more than 400M image-text pairs for training. The inefficiency can be \textit{partially} attributed to the fact that the image-text pairs are noisy. To address this, we propose OTTER (Optimal TransporT distillation for Efficient zero-shot Recognition), which uses online entropic optimal transport to find a soft image-text match as labels for contrastive learning. Based on pretrained image and text encoders, models trained with OTTER achieve strong performance with only 3M image text pairs. Compared with InfoNCE loss, label smoothing, and knowledge distillation, OTTER consistently outperforms these baselines in zero-shot evaluation on Google Open Images (19,958 classes) and multi-labeled ImageNet 10K (10032 classes) from Tencent ML-Images. Over 42 evaluations on 7 different dataset/architecture settings x 6 metrics, OTTER outperforms (32) or ties (2) all baselines in 34 of them. Our source code is open sourced at https://github.com/facebookresearch/OTTER.

****

Meaningfully Explaining Model Mistakes Using Conceptual Counterfactuals

Abubakar Abid,Mert Yuksekgonul,James Zou

Understanding and explaining the mistakes made by trained models is critical to many machine learning objectives, such as improving robustness, addressing concept drift, and mitigating biases. However, this is often an ad hoc process that involves manually looking at the model's mistakes on many test samples and guessing at the underlying reasons for those incorrect predictions. In this paper, we propose a systematic approach, \textit{conceptual counterfactual explanations} (CCE), that explains why a classifier makes a mistake on a particular test sample(s) in terms of human-understandable concepts (e.g. this zebra is misclassified as a dog because of faint \emph{stripes}). We base CCE on two prior ideas: counterfactual explanations and concept activation vectors, and validate our approach on well-known pretrained models, showing that it explains the models' mistakes meaningfully. In addition, for new models trained on data with spurious correlations, CCE accurately identifies the spurious correlation as the cause of model mistakes from a single misclassified test sample. On two challenging medical applications, CCE generated useful insights, confirmed by clinicians, into biases and mistakes the model makes in real-world settings. The code for CCE is publicly available and can easily be applied to explain mistakes in new models.

****

Robust Losses for Learning Value Functions

Andrew Patterson,Victor Liao,Martha White

Most value function learning algorithms in reinforcement learning are based on the mean squared (projected) Bellman error. However, squared errors are known to be sensitive to outliers, both skewing the solution of the objective and resulting in high-magnitude and high-variance gradients. Typical strategies to control these high-magnitude updates in RL involve clipping gradients, clipping rewards, rescaling rewards, and clipping errors. Clipping errors is related to using robust losses, like the Huber loss, but as yet no work explicitly formalizes and derives value learning algorithms with robust losses. In this work, we build on recent insights reformulating squared Bellman errors as a saddlepoint optimization problem, and propose a saddlepoint reformulation for a Huber Bellman error and

Absolute Bellman error. We show that the resulting solutions have significantly lower error for certain problems and are otherwise comparable, in terms of both absolute and squared value error. We show that the resulting gradient-based algorithms are more robust, for both prediction and control, with less stepsize sensitivity.

****

Feudal Reinforcement Learning by Reading Manuals

Kai Wang,Zhonghao Wang,Mo Yu,Humphrey Shi

Reading to act is a prevalent but challenging task that requires the ability to follow a concise language instruction in an environment, with the help of textual knowledge about the environment. Previous works face the semantic mismatch between the low-level actions and the high-level language descriptions and require the human-designed curriculum to work properly. In this paper, we present a Feudal Reinforcement Learning (FRL) model consisting of a manager agent and a worker agent. The manager agent is a multi-hop planner, which deals with high-level abstract information and generates a series of sub-goals. The worker agent deals with the low-level perceptions and actions to achieve the sub-goals one by one. Our FRL framework effectively alleviates the mismatching between the text-level inference and the low-level perceptions and actions; and is general to various forms of environments, instructions and manuals. Our multi-hop planner contributes to the framework by further boosting the challenging tasks where multi-step reasoning from the texts is critical to achieving the instructed goals. We showcase our approach achieves competitive performance on two challenging tasks, Read to Fight Monsters (RTFM) and Messenger, without human-designed curriculum learning.

****

Propagating Distributions through Neural Networks

Felix Petersen,Christian Borgelt,Mikhail Yurochkin,Hilde Kuehne,Oliver Deussen

We propose a new approach to propagating probability distributions through neural networks. To handle non-linearities, we use local linearization and show this to be an optimal approximation in terms of total variation for ReLUs. We demonstrate the advantages of our method over the moment matching approach popularized in prior works. In addition, we formulate new loss functions for training neural networks based on distributions. To demonstrate the utility of propagating distributions, we apply it to quantifying prediction uncertainties. In regression tasks we obtain calibrated confidence intervals, and in a classification setting we improve selective prediction on out-of-distribution data. We also show empirically that training with our uncertainty aware losses improve robustness to random and adversarial noise.

****

A Principled Permutation Invariant Approach to Mean-Field Multi-Agent Reinforcement Learning

Yan Li,Lingxiao Wang,Jiachen Yang,Ethan Wang,Zhaoran Wang,Tuo Zhao,Hongyuan Zha

Multi-agent reinforcement learning (MARL) becomes more challenging in the presence of more agents, as the capacity of the joint state and action spaces grows exponentially in the number of agents. To address such a challenge of scale, we identify a class of cooperative MARL problems with permutation invariance, and formulate it as mean-field Markov decision processes (MDP). To exploit the permutation invariance therein, we propose the mean-field proximal policy optimization (MF-PPO) algorithm, at the core of which is a permutation- invariant actor-critic neural architecture. We prove that MF-PPO attains the globally optimal policy at a sublinear rate of convergence. Moreover, its sample complexity is independent of the number of agents. We validate the theoretical advantages of MF-PPO with numerical experiments in the multi-agent particle environment (MPE). In particular, we show that the inductive bias introduced by the permutation-invariant neural architecture enables MF-PPO to outperform existing competitors with a smaller number of model parameters, which is the key to its generalization performance.

****

Where do Models go Wrong? Parameter-Space Saliency Maps for Explainability

Roman Levin,Manli Shu,Eitan Borgnia,Furong Huang,Micah Goldblum,Tom Goldstein

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 pruning 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.

****

A Statistical Framework for Efficient Out of Distribution Detection in Deep Neural Networks

Matan Haroush,Tzviel Frostig,Ruth Heller,Daniel Soudry

Background.

Commonly, Deep Neural Networks (DNNs) generalize well on samples drawn from a distribution similar to that of the training set. However, DNNs' predictions are brittle and unreliable when the test samples are drawn from a dissimilar distribution.

This is a major concern for deployment in real-world applications, where such behavior may come at a considerable cost, such as industrial production lines, autonomous vehicles, or healthcare applications.

Contributions.

We frame Out Of Distribution (OOD) detection in DNNs as a statistical hypothesis testing problem. Tests generated within our proposed framework combine evidence from the entire network.

Unlike previous OOD detection heuristics, this framework returns a p-value for each test sample. It is guaranteed to maintain the Type I Error (T1E - incorrectly predicting OOD for an actual in-distribution sample) for test data. Moreover, this allows to combine several detectors while maintaining the T1E.

Building on this framework, we suggest a novel OOD procedure based on low-order statistics. Our method achieves comparable or better results than state-of-the-art methods on well-accepted OOD benchmarks, without retraining the network parameters or assuming prior knowledge on the test distribution --- and at a fraction of the computational cost.

****

FedBABU: Toward Enhanced Representation for Federated Image Classification

Jaehoon Oh,SangMook Kim,Se-Young Yun

Federated learning has evolved to improve a single global model under data heterogeneity (as a curse) or to develop multiple personalized models using data heterogeneity (as a blessing). However, little research has considered both directions simultaneously. In this paper, we first investigate the relationship between them by analyzing Federated Averaging at the client level and determine that a better federated global model performance does not constantly improve personalization. To elucidate the cause of this personalization performance degradation problem, we decompose the entire network into the body (extractor), which is related to universality, and the head (classifier), which is related to personalization. We then point out that this problem stems from training the head. Based on this observation, we propose a novel federated learning algorithm, coined FedBABU, which only updates the body of the model during federated training (i.e., the head is randomly initialized and never updated), and the head is fine-tuned for p ersonalization during the evaluation process. Extensive experiments show consistent performance improvements and an efficient personalization of FedBABU. The code is available at https://github.com/jhoon-oh/FedBABU.

****

Should I Run Offline Reinforcement Learning or Behavioral Cloning?

Aviral Kumar,Joey Hong,Anikait Singh,Sergey Levine

Offline reinforcement learning (RL) algorithms can acquire effective policies by utilizing only previously collected experience, without any online interaction. While it is widely understood that offline RL is able to extract good policies even from highly suboptimal data, in practice offline RL is often used with data that resembles demonstrations. In this case, one can also use behavioral cloning (BC) algorithms, which mimic a subset of the dataset via supervised learning. It seems natural to ask: When should we prefer offline RL over BC? In this paper, our goal is to characterize environments and dataset compositions where offline RL leads to better performance than BC. In particular, we characterize the properties of environments that allow offline RL methods to perform better than BC methods even when only provided with expert data. Additionally, we show that policies trained on suboptimal data that is sufficiently noisy can attain better performance than even BC algorithms with expert data, especially on long-horizon problems. We validate our theoretical results via extensive experiments on both diagnostic and high-dimensional domains including robot manipulation, maze navigation and Atari games, when learning from a variety of data sources. We observe that modern offline RL methods trained on suboptimal, noisy data in sparse reward domains outperform cloning the expert data in several practical problems.

****

MT3: Multi-Task Multitrack Music Transcription

Joshua P Gardner,Ian Simon,Ethan Manilow,Curtis Hawthorne,Jesse Engel

Automatic Music Transcription (AMT), inferring musical notes from raw audio, is a challenging task at the core of music understanding. Unlike Automatic Speech Recognition (ASR), which typically focuses on the words of a single speaker, AMT often requires transcribing multiple instruments simultaneously, all while preserving fine-scale pitch and timing information. Further, many AMT datasets are ``low-resource'', as even expert musicians find music transcription difficult and time-consuming. Thus, prior work has focused on task-specific architectures, tailored to the individual instruments of each task. In this work, motivated by the promising results of sequence-to-sequence transfer learning for low-resource Natural Language Processing (NLP), we demonstrate that a general-purpose Transformer model can perform multi-task AMT, jointly transcribing arbitrary combinations of musical instruments across several transcription datasets. We show this unified training framework achieves high-quality transcription results across a range of datasets, dramatically improving performance for low-resource instruments (such as guitar), while preserving strong performance for abundant instruments (such as piano). Finally, by expanding the scope of AMT, we expose the need for more consistent evaluation metrics and better dataset alignment, and provide a strong baseline for this new direction of multi-task AMT.

****

Learning State Representations via Retracing in Reinforcement Learning

Changmin Yu,Dong Li,Jianye HAO,Jun Wang,Neil Burgess

We propose learning via retracing, a novel self-supervised approach for learning the state representation (and the associated dynamics model) for reinforcement learning tasks. In addition to the predictive (reconstruction) supervision in the forward direction, we propose to include "retraced" transitions for representation/model learning, by enforcing the cycle-consistency constraint between the original and retraced states, hence improve upon the sample efficiency of learning. Moreover, learning via retracing explicitly propagates information about future transitions backward for inferring previous states, thus facilitates stronger representation learning for the downstream reinforcement learning tasks. We introduce Cycle-Consistency World Model (CCWM), a concrete model-based instantiation of learning via retracing. Additionally we propose a novel adaptive "truncation" mechanism for counteracting the negative impacts brought by "irreversible" tr ansitions such that learning via retracing can be maximally effective. Through extensive empirical studies on visual-based continuous control benchmarks, we demonstrate that CCWM achieves state-of-the-art performance in terms of sample efficiency and asymptotic performance, whilst exhibiting behaviours that are indicative of stronger representation learning.

****

Novelty detection using ensembles with regularized disagreement

Alexandru Tifrea,Eric Petru Stavarache,Fanny Yang

Despite their excellent performance on in-distribution (ID) data, deep neural networks often confidently predict on out-of-distribution (OOD) samples that come from novel classes instead of flagging them for expert evaluation. Even though conventional OOD detection algorithms can distinguish far OOD samples, current methods that can identify near OOD samples require training with labeled data that is very similar to these near OOD samples. In turn, we develop a new ensemble-based procedure for \emph{semi-supervised novelty detection} (SSND) that only utilizes a mixture of unlabeled ID and OOD samples to achieve good detection performance on near OOD data. It crucially relies on regularization to promote diversity on the OOD data while preserving agreement on ID data. Extensive comparisons of our approach to state-of-the-art SSND methods on standard image data sets (SVHN/CIFAR-10/CIFAR-100) and medical image data sets reveal significant gains with negligible increase in computational cost.

****

Open-World Semi-Supervised Learning

Kaidi Cao,Maria Brbic,Jure Leskovec

A fundamental limitation of applying semi-supervised learning in real-world settings is the assumption that unlabeled test data contains only classes previously encountered in the labeled training data. However, this assumption rarely holds for data in-the-wild, where instances belonging to novel classes may appear at testing time. Here, we introduce a novel open-world semi-supervised learning setting that formalizes the notion that novel classes may appear in the unlabeled test data. In this novel setting, the goal is to solve the class distribution mismatch problem between labeled and unlabeled data, where at the test time every input instance either needs to be classified into one of the existing classes or a new unseen class needs to be initialized and the instance assigned to it. To tackle this challenging problem, we propose ORCA, an end-to-end approach that assigns instances to previously seen classes or forms novel classes by grouping similar instances without assuming any prior knowledge. The key idea in ORCA is to utilize uncertainty adaptive margin to circumvent the bias towards seen classes caused by learning seen classes faster than the novel classes. In this way, ORCA gradually increases the discriminability of the model during the training and reduces the gap between intra-class variance of seen with respect to novel classes. Extensive experiments on image classification datasets and a single-cell dataset demonstrate that ORCA consistently outperforms alternative baselines, achieving 25% improvement on seen and 96% improvement on novel classes of the ImageNet dataset.

****

Guiding Transformers to Process in Steps

Simas Sakenis,Stuart Shieber

Neural networks have matched or surpassed human abilities in many tasks that humans solve quickly and unconsciously, i.e., via Kahneman's “System 1”, but have not been as successful when applied to “System 2” tasks that involve conscious multi-step reasoning. In this work, we argue that the kind of training that works for System 1 tasks is not sufficient for System 2 tasks, propose an alternative, and empirically demonstrate its effectiveness. Specifically, while learning a direct mapping from inputs to outputs is feasible for System 1 tasks, we argue that algorithmic System 2 tasks can only be solved by learning a mapping from inputs to outputs through a series of intermediate steps. We first show that by using enough intermediate steps a 1-layer 1-head Transformer can in principle compute any finite function, proving the generality of the approach. We then show empi rically that a 1-layer 1-head Transformer cannot learn to compute the sum of binary numbers directly from the inputs, but is able to compute the sum when trained to first generate a series of intermediate results. This demonstrates, at a small scale, how a fixed-size neural network can lack the expressivity to encode the direct input-output mapping for an algorithmic task and yet be fully capable of computing the outputs through intermediate steps. Finally, we show that a Frozen Pretrained Transformer is able to learn binary addition when trained to compute the carry bits before the sum, while it fails to learn the task without using intermediates. These results indicate that explicitly guiding the neural networks through the intermediate computations can be an effective approach for tackling algorithmic tasks.

****

Introspective Learning : A Two-Stage approach for Inference in Neural Networks
Mohit Prabhushankar,Ghassan AlRegib

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. We perceptually visualize the 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 and out-of-distribution detection. Finally, we ground the proposed machine introspection to human introspection in the application of image quality assessment.

****

C5T5: Controllable Generation of Organic Molecules with Transformers
Daniel Rothchild,Alex Tamkin,Julie Yu,Ujval Misra,Joseph E. Gonzalez

Methods for designing organic materials with desired properties have high potential impact across fields such as medicine, renewable energy, petrochemical engineering, and agriculture. However, using generative models for this task is difficult because candidate compounds must satisfy many constraints, including synthetic accessibility, intellectual property attributes, ``chemical beauty'' (Bickerton et al., 2020), and other considerations that are intuitive to domain experts but can be challenging to quantify. We propose C5T5, a novel self-supervised pretraining method that works in tandem with domain experts by making zero-shot select-and-replace edits, altering organic substances towards desired property values. C5T5 operates on IUPAC names---a standardized molecular representation that intuitively encodes rich structural information for organic chemists but that has been largely ignored by the ML community. Our technique requires no edited molecule pairs to train and only a rough estimate of molecular properties, and it has the potential to model long-range dependencies and symmetric molecular structures more easily than graph-based methods. We demonstrate C5T5's effectiveness on four physical properties relevant for drug discovery, showing that it learns successful and chemically intuitive strategies for altering molecules towards desired property values.

****

Compressed-VFL: Communication-Efficient Learning with Vertically Partitioned Data
Timothy Castiglia,Anirban Das,Shiqiang Wang,Stacy Patterson

We propose Compressed Vertical Federated Learning (C-VFL) for communication-efficient training on vertically partitioned data. In C-VFL, a server and multiple parties collaboratively train a model on their respective features utilizing several local iterations and sharing compressed intermediate results periodically. Our work provides the first theoretical analysis of the effect message compressio n has on distributed training over vertically partitioned data. We prove convergence of non-convex objectives to a fixed point at a rate of $O(\frac{1}{\sqrt{T}})$ when the compression error is bounded over the course of training. We provide specific requirements for convergence with common compression techniques, such as quantization and top-k sparsification. Finally, we experimentally show compression can reduce communication by over 90% without a significant decrease in accuracy over VFL without compression.

****

Causal Reinforcement Learning using Observational and Interventional Data
Maxime Gasse,Damien GRASSET,Guillaume Gaudron,Pierre-Yves Oudeyer

Learning efficiently a causal model of the environment is a key challenge of model-based RL agents operating in POMDPs. We consider here a scenario where the learning agent has the ability to collect online experiences through direct interactions with the environment (interventional data), but also has access to a large collection of offline experiences, obtained by observing another agent interacting with the environment (observational data). A key ingredient, which makes this situation non-trivial, is that we allow the observed agent to act based on privileged information, hidden from the learning agent. We then ask the following questions: can the online and offline experiences be safely combined for learning a causal transition model ? And can we expect the offline experiences to improve the agent's performances ? To answer these, first we bridge the fields of reinforcement learning and causality, by importing ideas from the well-established causal framework of do-calculus, and expressing model-based reinforcement learning as a causal inference problem. Second, we propose a general yet simple methodology for safely leveraging offline data during learning. In a nutshell, our method relies on learning a latent-based causal transition model that explains both the interventional and observational regimes, and then inferring the standard POMDP transition model via deconfounding using the recovered latent variable. We prove our method is correct and efficient in the sense that it attains better generalization guarantees due to the offline data (in the asymptotic case), and we assess its effectiveness empirically on a series of synthetic toy problems.

****

Data-Efficient Graph Grammar Learning for Molecular Generation
Minghao Guo,Veronika Thost,Beichen Li,Payel Das,Jie Chen,Wojciech Matusik

The problem of molecular generation has received significant attention recently. Existing methods are typically based on deep neural networks and require training on large datasets with tens of thousands of samples. In practice, however, the size of class-specific chemical datasets is usually limited (e.g., dozens of samples) due to labor-intensive experimentation and data collection. Another major challenge is to generate only physically synthesizable molecules. This is a non-trivial task for neural network-based generative models since the relevant chemical knowledge can only be extracted and generalized from the limited training data. In this work, we propose a data-efficient generative model that can be learned from datasets with orders of magnitude smaller sizes than common benchmarks. At the heart of this method is a learnable graph grammar that generates molecules from a sequence of production rules. Without any human assistance, these production rules are automatically constructed from training data. Furthermore, additional chemical knowledge can be incorporated into the model by further grammar optimization. Our learned graph grammar yields state-of-the-art results on generating high-quality molecules for three monomer datasets that contain only ${\sim}20$ samples each. Our approach also achieves remarkable performance in a challenging polymer generation task with $only$ 117 training samples and is competitive against existing methods using 81k data points.

****

Data Sharing without Rewards in Multi-Task Offline Reinforcement Learning
Tianhe Yu,Aviral Kumar,Yevgen Chebotar,Chelsea Finn,Sergey Levine,Karol Hausman

Offline reinforcement learning (RL) bears the promise to learn effective control policies from static datasets but is thus far unable to learn from large databases of heterogeneous experience. The multi-task version of offline RL enables th

e possibility of learning a single policy that can tackle multiple tasks and allows the algorithm to share offline data across tasks. Recent works indicate that sharing data between tasks can be highly beneficial in multi-task learning. However, these benefits come at a cost -- for data to be shared between tasks, each transition must be annotated with reward labels corresponding to other tasks. This is particularly expensive and unscalable, since the manual effort in annotating reward grows quadratically with the number of tasks. Can we retain the benefits of data sharing without requiring reward relabeling for every task pair? In this paper, we show that, perhaps surprisingly, under a binary-reward assumption, simply utilizing data from other tasks with constant reward labels can not only provide a substantial improvement over only using the single-task data and previously proposed success classifiers, but it can also reach comparable performance to baselines that take advantage of the oracle multi-task reward information. We also show that this performance can be further improved by selectively deciding which transitions to share, again without introducing any additional models or classifiers. We discuss how these approaches relate to each other and baseline strategies under various assumptions on the dataset. Our empirical results show that it leads to improved performance across a range of different multi-task offline RL scenarios, including robotic manipulation from visual inputs and ant-maze navigation.
****
Evading Adversarial Example Detection Defenses with Orthogonal Projected Gradient Descent
Oliver Bryniarski,Nabeel Hingun,Pedro Pachuca,Vincent Wang,Nicholas Carlini
Evading adversarial example detection defenses requires finding adversarial examples that must simultaneously (a) be misclassified by the model and (b) be detected as non-adversarial. We find that existing attacks that attempt to satisfy multiple simultaneous constraints often over-optimize against one constraint at the cost of satisfying another. We introduce Selective Projected Gradient Descent and Orthogonal Projected Gradient Descent, improved attack techniques to generate adversarial examples that avoid this problem by orthogonalizing the gradients when running standard gradient-based attacks. We use our technique to evade four state-of-the-art detection defenses, reducing their accuracy to 0% while maintaining a 0% detection rate.
****
Dominant Datapoints and the Block Structure Phenomenon in Neural Network Hidden Representations
Thao Nguyen,Maithra Raghu,Simon Kornblith
Recent work has uncovered a striking phenomenon in large-capacity neural networks: they contain blocks of contiguous hidden layers with highly similar representations. This block structure has two seemingly contradictory properties: on the one hand, its constituent layers have highly similar dominant first principal components (PCs), but on the other hand, their representations, and their common first PC, are highly dissimilar across different random seeds. Our work seeks to reconcile these discrepant properties by investigating the origin of the block structure in relation to the data and training methods. By analyzing properties of the dominant PCs, we find that the block structure arises from dominant datapoints — a small group of examples that share similar image statistics (e.g. background color). However, the set of dominant datapoints, and the precise shared image statistic, can vary across random seeds. Thus, the block structure reflects meaningful dataset statistics, but is simultaneously unique to each model. Through studying hidden layer activations and creating synthetic datapoints, we demonstrate that these simple image statistics dominate the representational geometry of the layers inside the block structure. We also explore how the phenomenon evolves through training, finding that the block structure takes shape early in training, but the underlying representations and the corresponding dominant datapoints continue to change substantially. Finally, we study the interplay between the block structure and different training mechanisms, introducing a targeted intervention to eliminate the block structure, as well as examining the effects of pretraining and Shake-Shake regularization.

****

On Optimal Early Stopping: Overparametrization versus Underparametrization
Ruoqi Shen,Liyao Gao,Yian Ma

Early stopping is a simple and widely used method to prevent over-training neural networks. We develop theoretical results to reveal the relationship between optimal early stopping time and model dimension as well as sample size of the dataset for certain linear regression models. Our results demonstrate two very different behaviors when the model dimension exceeds the number of features versus the opposite scenario. While most previous works on linear models focus on the latter setting, we observe that in common deep learning tasks, the dimension of the model often exceeds the number of features arising from data. We demonstrate experimentally that our theoretical results on optimal early stopping time corresponds to the training process of deep neural network. Moreover, we study the effect of early stopping on generalization and demonstrate that optimal early stopping can help mitigate ''descent'' in various settings.

****

Reducing Excessive Margin to Achieve a Better Accuracy vs. Robustness Trade-off
Rahul Rade,Seyed-Mohsen Moosavi-Dezfooli

While adversarial training has become the de facto approach for training robust classifiers, it leads to a drop in accuracy. This has led to prior works postulating that accuracy is inherently at odds with robustness. Yet, the phenomenon remains inexplicable. In this paper, we closely examine the changes induced in the decision boundary of a deep network during adversarial training. We find that adversarial training leads to unwarranted increase in the margin along certain adversarial directions, thereby hurting accuracy. Motivated by this observation, we present a novel algorithm, called Helper-based Adversarial Training (HAT), to reduce this effect by incorporating additional wrongly labelled examples during training. Our proposed method provides a notable improvement in accuracy without compromising robustness. It achieves a better trade-off between accuracy and robustness in comparison to existing defenses. Code is available at https://github.com/imrahulr/hat.

****

Expressivity of Emergent Languages is a Trade-off between Contextual Complexity and Unpredictability
Shangmin Guo,Yi Ren,Kory Wallace Mathewson,Simon Kirby,Stefano V Albrecht,Kenny Smith

Researchers are using deep learning models to explore the emergence of language in various language games, where agents interact and develop an emergent language to solve tasks. We focus on the factors that determine the expressivity of emergent languages, which reflects the amount of information about input spaces those languages are capable of encoding. We measure the expressivity of emergent languages based on the generalisation performance across different games, and demonstrate that the expressivity of emergent languages is a trade-off between the complexity and unpredictability of the context those languages emerged from. Another contribution of this work is the discovery of message type collapse, i.e. the number of unique messages is lower than that of inputs. We also show that using the contrastive loss proposed by Chen et al. (2020) can alleviate this problem.

****

Resmax: An Alternative Soft-Greedy Operator for Reinforcement Learning
Erfan Miahi,Revan MacQueen,Alex Ayoub,Abbas Masoumzadeh,Martha White

Soft-greedy operators, namely ε-greedy and softmax, remain a common choice to induce a basic level of exploration for action-value methods in reinforcement learning. These operators, however, have a few critical limitations. In this work, we investigate a simple soft-greedy operator, which we call resmax, that takes actions proportionally to their suboptimality gap: the residual to the estimated maximal value. It is simple to use and ensures coverage of the state-space like ε-greedy, but focuses exploration more on potentially promising actions like softmax. Further, it does not concentrate probability as quickly as softmax, and so better avoids overemphasizing sub-optimal actions that a ppear high-valued during learning. Additionally, we prove it is a non-expansion for any fixed exploration hyperparameter, unlike the softmax policy which requires a state-action specific temperature to obtain a non-expansion (called mellowmax). We empirically validate that resmax is comparable to or outperforms ε-greedy and softmax across a variety of environments in tabular and deep RL.

****

Fast AdvProp

Jieru Mei,Yucheng Han,Yutong Bai,Yixiao Zhang,Yingwei Li,Xianhang Li,Alan Yuille,Cihang Xie

Adversarial Propagation (AdvProp) is an effective way to improve recognition models, leveraging adversarial examples. Nonetheless, AdvProp suffers from the extremely slow training speed, mainly because: a) extra forward and backward passes are required for generating adversarial examples; b) both original samples and their adversarial counterparts are used for training (i.e., 2X data). In this paper, we introduce Fast AdvProp, which aggressively revamps AdvProp's costly training components, rendering the method nearly as cheap as the vanilla training. Specifically, our modifications in Fast AdvProp are guided by the hypothesis that disentangled learning with adversarial examples is the key for performance improvements, while other training recipes (e.g., paired clean and adversarial training samples, multi-step adversarial attackers) could be largely simplified.

Our empirical results show that, compared to the vanilla training baseline, Fast AdvProp is able to further model performance on a spectrum of visual benchmarks, without incurring extra training cost. Additionally, our ablations find Fast AdvProp scales better if larger models are used, is compatible with existing data augmentation methods (i.e., Mixup and CutMix), and can be easily adapted to other recognition tasks like object detection. The code is available here: https://github.com/meijieru/fast_advprop.

****

Rethinking the limiting dynamics of SGD: modified loss, phase space oscillations, and anomalous diffusion

Daniel Kunin,Javier Sagastuy-Brena,Lauren Gillespie,Eshed Margalit,Hidenori Tanaka,Surya Ganguli,Daniel LK Yamins

In this work we explore the limiting dynamics of deep neural networks trained with stochastic gradient descent (SGD). We find empirically that long after performance has converged, networks continue to move through parameter space by a process of anomalous diffusion in which distance travelled grows as a power law in the number of gradient updates with a nontrivial exponent. We reveal an intricate interaction between the hyperparameters of optimization, the structure in the gradient noise, and the Hessian matrix at the end of training that explains this anomalous diffusion. To build this understanding, we first derive a continuous-time model for SGD with finite learning rates and batch sizes as an underdamped Langevin equation. We study this equation in the setting of linear regression, where we can derive exact, analytic expressions for the phase space dynamics of the parameters and their instantaneous velocities from initialization to stationarity. Using the Fokker-Planck equation, we show that the key ingredient driving these dynamics is not the original training loss, but rather the combination of a modified loss, which implicitly regularizes the velocity, and probability currents, which cause oscillations in phase space. We identify qualitative and quantitative predictions of this theory in the dynamics of a ResNet-18 model trained on ImageNet. Through the lens of statistical physics, we uncover a mechanistic origin for the anomalous limiting dynamics of deep neural networks trained with SGD.

****

Adversarial Style Transfer for Robust Policy Optimization in Reinforcement Learning

Md Masudur Rahman,Yexiang Xue

This paper proposes an algorithm that aims to improve generalization for reinforcement learning agents by removing overfitting to confounding features. Our appr oach consists of a max-min game theoretic objective. A generator transfers the style of observation during reinforcement learning. An additional goal of the generator is to perturb the observation, which maximizes the agent's probability of taking a different action. In contrast, a policy network updates its parameters to minimize the effect of such perturbations, thus staying robust while maximizing the expected future reward. Based on this setup, we propose a practical deep reinforcement learning algorithm, Adversarial Robust Policy Optimization (ARPO), to find an optimal policy that generalizes to unseen environments. We evaluate our approach on visually enriched and diverse Procgen benchmarks. Empirically, we observed that our agent ARPO performs better in generalization and sample efficiency than a few state-of-the-art algorithms.

****

Poisoning and Backdooring Contrastive Learning

Nicholas Carlini,Andreas Terzis

Multimodal contrastive learning methods like CLIP train on noisy and uncurated training datasets. This is cheaper than labeling datasets manually, and even improves out-of-distribution robustness. We show that this practice makes backdoor and poisoning attacks a significant threat. By poisoning just 0.01% of a dataset (e.g., just 300 images of the 3 million-example Conceptual Captions dataset), we can cause the model to misclassify test images by overlaying a small patch. Targeted poisoning attacks, whereby the model misclassifies a particular test input with an adversarially-desired label, are even easier requiring control of 0.0001% of the dataset (e.g., just three out of the 3 million images). Our attacks call into question whether training on noisy and uncurated Internet scrapes is desirable.

****

Domain Invariant Adversarial Learning

Matan Levi,Idan Attias,Aryeh Kontorovich

The phenomenon of adversarial examples illustrates one of the most basic vulnerabilities of deep neural networks. Among the variety of techniques introduced to surmount this inherent weakness, adversarial training has emerged as the most effective strategy to achieve robustness. Typically, this is achieved by balancing robust and natural objectives. In this work, we aim to further reduce the trade-off between robust and standard accuracy by enforcing a domain-invariant feature representation. We present a new adversarial training method, Domain Invariant Adversarial Learning (DIAL), which learns a feature representation which is both robust and domain invariant. DIAL uses a variant of Domain Adversarial Neural Network (DANN) on the natural domain and its corresponding adversarial domain. In a case where the source domain consists of natural examples and the target domain is the adversarially perturbed examples, our method learns a feature representation constrained not to discriminate between the natural and adversarial examples, and can therefore achieve a more robust representation. Our experiments indicate that our method improves both robustness and standard accuracy, when compared to other state-of-the-art adversarial training methods.

****

Multi-Tailed, Multi-Headed, Spatial Dynamic Memory refined Text-to-Image Synthesis

Amrit Diggavi Seshadri,Balaraman Ravindran

Synthesizing high-quality, realistic images from text-descriptions is a challenging task, and current methods synthesize images from text in a multi-stage manner, typically by first generating a rough initial image and then refining image details at subsequent stages. However, existing methods that follow this paradigm suffer from three important limitations. Firstly, they synthesize initial images without attempting to separate image attributes at a word-level. As a result, object attributes of initial images (that provide a basis for subsequent refinement) are inherently entangled and ambiguous in nature. Secondly, by using common text-representations for all regions, current methods prevent us from interpreting text in fundamentally different ways at different parts of an image. Different image regions are therefore only allowed to assimilate the same type of information from text at each refinement stage. Finally, current methods generate ref inement features only once at each refinement stage and attempt to address all image aspects in a single shot. This single-shot refinement limits the precision with which each refinement stage can learn to improve the prior image. Our proposed method introduces three novel components to address these shortcomings: (1) An initial generation stage that explicitly generates separate sets of image features for each word n-gram. (2) A spatial dynamic memory module for refinement of images. (3) An iterative multi-headed mechanism to make it easier to improve upon multiple image aspects. Experimental results demonstrate that our Multi-Headed Spatial Dynamic Memory image refinement with our Multi-Tailed Word-level Initial Generation (MSMT-GAN) performs favourably against the previous state of the art on the CUB and COCO datasets.

****

Triangle and Four Cycle Counting with Predictions in Graph Streams

Justin Y Chen,Talya Eden,Piotr Indyk,Honghao Lin,Shyam Narayanan,Ronitt Rubinfeld,Sandeep Silwal,Tal Wagner,David Woodruff,Michael Zhang

We propose data-driven one-pass streaming algorithms for estimating the number of triangles and four cycles, two fundamental problems in graph analytics that are widely studied in the graph data stream literature. Recently, Hsu et al. (2019) and Jiang et al. (2020) applied machine learning techniques in other data stream problems, using a trained oracle that can predict certain properties of the stream elements to improve on prior “classical” algorithms that did not use oracles. In this paper, we explore the power of a “heavy edge” oracle in multiple graph edge streaming models. In the adjacency list model, we present a one-pass triangle counting algorithm improving upon the previous space upper bounds without such an oracle. In the arbitrary order model, we present algorithms for both triangle and four cycle estimation with fewer passes and the same space complexity as in previous algorithms, and we show several of these bounds are optimal. We analyze our algorithms under several noise models, showing that the algorithms perform well even when the oracle errs. Our methodology expands upon prior work on “classical” streaming algorithms, as previous multi-pass and random order streaming algorithms can be seen as special cases of our algorithms, where the first pass or random order was used to implement the heavy edge oracle. Lastly, our experiments demonstrate advantages of the proposed method compared to state-of-the-art streaming algorithms.

****

Learning Stable Classifiers by Transferring Unstable Features

Yujia Bao,Shiyu Chang,Regina Barzilay

While unbiased machine learning models are essential for many applications, bias is a human-defined concept that can vary across tasks. Given only input-label pairs, algorithms may lack sufficient information to distinguish stable (causal) features from unstable (spurious) features. However, related tasks often share similar biases -- an observation we may leverage to develop stable classifiers in the transfer setting. In this work, we explicitly inform the target classifier about unstable features in the source tasks. Specifically, we derive a representation that encodes the unstable features by contrasting different data environments in the source task. We achieve robustness by clustering data of the target task according to this representation and minimizing the worst-case risk across these clusters. We evaluate our method on both text and image classifications. Empirical results demonstrate that our algorithm is able to maintain robustness on the target task, outperforming the best baseline by 22.9% in absolute accuracy across 12 transfer settings. Our code and data will be publicly available.

****

Connectivity Matters: Neural Network Pruning Through the Lens of Effective Sparsity

Artem M Vysogorets,Julia Kempe

Neural network pruning is a fruitful area of research with surging interest in high sparsity regimes. Benchmarking in this domain heavily relies on faithful representation of the sparsity of subnetworks, which has been traditionally computed as the fraction of removed connections (direct sparsity). This definition, however, fails to recognize unpruned parameters that detached from input or output layers of underlying subnetworks, potentially underestimating actual effective sparsity: the fraction of inactivated connections. While this effect might be negligible for moderately pruned networks (up to $10\times-100\times$ compression rates), we find that it plays an increasing role for thinner subnetworks, greatly distorting comparison between different pruning algorithms. For example, we show that effective compression of a randomly pruned LeNet-300-100 can be orders of magnitude larger than its direct counterpart, while no discrepancy is ever observed when using SynFlow for pruning (Tanaka et al., 2020). In this work, we adopt the lens of effective sparsity to reevaluate several recent pruning algorithms on common benchmark architectures (e.g., LeNet-300-100, VGG-19, ResNet-18) and discover that their absolute and relative performance changes dramatically in this new, and as we argue, more appropriate framework. To aim for effective, rather than direct, sparsity, we develop a low-cost extension to most pruning algorithms. Further, equipped with effective sparsity as a reference frame, we partially reconfirm that random pruning with appropriate sparsity allocation across layers performs as well or better than more sophisticated algorithms for pruning at initialization (Su et al., 2020). In response to this observation, using a simple analogy of pressure distribution in coupled cylinders from thermodynamics, we design novel layerwise sparsity quotas that outperform all existing baselines in the context of random pruning.

****

Is Fairness Only Metric Deep? Evaluating and Addressing Subgroup Gaps in Deep Metric Learning
Natalie Dullerud,Karsten Roth,Kimia Hamidieh,Nicolas Papernot,Marzyeh Ghassemi
Deep metric learning (DML) enables learning with less supervision through its emphasis on the similarity structure of representations. There has been much work on improving generalization of DML in settings like zero-shot retrieval, but little is known about its implications for fairness. In this paper, we are the first to evaluate state-of-the-art DML methods trained on imbalanced data, and to show the negative impact these representations have on minority subgroup performance when used for downstream tasks. In this work, we first define fairness in DML through an analysis of three properties of the representation space -- inter-class alignment, intra-class alignment, and uniformity -- and propose \textit{\textbf{finDML}}, the \textit{\textbf{f}}airness \textit{\textbf{i}}n \textit{\textbf{n}}on-balanced \textit{\textbf{DML}} benchmark to characterize representation fairness. Utilizing \textit{finDML}, we find bias in DML representations to propagate to common downstream classification tasks. Surprisingly, this bias is propagated even when training data in the downstream task is re-balanced. To address this problem, we present Partial Attribute De-correlation (\textit{\textbf{\pad}}) to disentangle feature representations from sensitive attributes and reduce performance gaps between subgroups in both embedding space and downstream metrics.

****

Zero-Shot Coordination via Semantic Relationships Between Actions and Observations
Mingwei Ma,Jizhou Liu,Samuel Sokota,Max Kleiman-Weiner,Jakob Nicolaus Foerster
An unaddressed challenge in zero-shot coordination is to take advantage of the semantic relationship between the features of an action and the features of observations. Humans take advantage of these relationships in highly intuitive ways. For instance in the absence of a shared-language, we might point to the object we desire or hold up fingers to indicate how many objects we want. To address this challenge, we investigate the effect of network architecture on the propensity of learning algorithms to make use of these relationships in human-compatible ways. We find that attention-based architectures that jointly process a featurized representation of the observation and the action, have a better inductive bias for exploiting semantic relationships for zero-shot coordination. Excitingly, in a set of diagnostic tasks, these agents produce highly human-compatible policies, without requiring the symmetry relationships of the problems to be hard-coded.

****

Does your graph need a confidence boost? Convergent boosted smoothing on graphs with tabular node features

Jiuhai Chen,Jonas Mueller,Vassilis N. Ioannidis,Soji Adeshina,Yangkun Wang,Tom Goldstein,David Wipf

Many practical modeling tasks require making predictions using tabular data composed of heterogeneous feature types (e.g., text-based, categorical, continuous, etc.). In this setting boosted decision trees and related ensembling techniques generally dominate real-world applications involving iid training/test sets. However, when there are relations between samples and the iid assumption is no longer reasonable, it remains unclear how to incorporate these dependencies within existing boosting pipelines. To this end, we propose a generalized framework for combining boosted trees and more general model ensembling techniques, with graph propagation layers that share node/sample information across edges connecting related samples. And unlike previous efforts to integrate graph-based models with boosting, our approach is anchored to a principled meta loss function such that provable convergence can be guaranteed under relatively mild assumptions. Across a variety of benchmarks involving non-iid graph data with tabular node features, our framework achieves comparable or superior performance.

****

PDQN - A Deep Reinforcement Learning Method for Planning with Long Delays: Optimization of Manufacturing Dispatching

David C Jenkins,René Arendt Sørensen,Vikramank Singh,Philip Kaminsky,Anil Aswani,Ramakrishna Akella

Scheduling is an important component in Semiconductor Manufacturing systems, where decisions must be made as to how to prioritize the use of finite machine resources to complete operations on parts in a timely manner. Traditionally, Operations Research methods have been used for simple, less complex systems. However, due to the complexity of this scheduling problem, simple dispatching rules such as Critical Ratio, and First-In-First-Out, are often used in practice in the industry for these more complex factories. This paper proposes a novel method based on Deep Reinforcement Learning for developing dynamic scheduling policies through interaction with simulated stochastic manufacturing systems. We experiment with simulated systems based on a complex Western Digital semiconductor plant. Our method builds upon DeepMind’s Deep Q-network, and predictron methods to create a novel algorithm, Predictron Deep Q-network, which utilizes a predictron model as a trained planning model to create training targets for a Deep Q-Network based policy. In recent years, Deep Reinforcement Learning methods have shown state of the art performance on sequential decision-making processes in complex games such as Go. Semiconductor manufacturing systems, however, provide significant additional challenges due to complex dynamics, stochastic transitions, and long time horizons with the associated delayed rewards. In addition, dynamic decision policies need to account for uncertainties such as machine downtimes. Experimental results demonstrate that, in our simulated environments, the Predictron Deep Q-network outperforms the Deep Q-network, Critical Ratio, and First-In-First-Out dispatching policies on the task of minimizing lateness of parts.

****

NodePiece: Compositional and Parameter-Efficient Representations of Large Knowledge Graphs

Mikhail Galkin,Etienne Denis,Jiapeng Wu,William L. Hamilton

Conventional representation learning algorithms for knowledge graphs (KG) map each entity to a unique embedding vector.

Such a shallow lookup results in a linear growth of memory consumption for storing the embedding matrix and incurs high computational costs of working with real-world KGs.

Drawing parallels with subword tokenization commonly used in NLP, we explore the landscape of more parameter-efficient node embedding strategies with possibly sublinear memory requirements.

To this end, we propose NodePiece, an anchor-based approach to learn a fixed-size entity vocabulary.

In NodePiece, a vocabulary of subword/sub-entity units is constructed from ancho r nodes in a graph with known relation types. Given such a fixed-size vocabulary, it is possible to bootstrap an encoding and embedding for any entity, including those unseen during training.

Experiments show that NodePiece performs competitively in node classification, link prediction, and relation prediction tasks retaining less than 10% of explicit nodes in a graph as anchors and often having 10x fewer parameters. To this end, we show that a NodePiece-enabled model outperforms existing shallow models on a large OGB WikiKG 2 graph having 70x fewer parameters.

****

Pix2seq: A Language Modeling Framework for Object Detection

Ting Chen,Saurabh Saxena,Lala Li,David J. Fleet,Geoffrey Hinton

We present Pix2Seq, a simple and generic framework for object detection. Unlike existing approaches that explicitly integrate prior knowledge about the task, we cast object detection as a language modeling task conditioned on the observed pixel inputs. Object descriptions (e.g., bounding boxes and class labels) are expressed as sequences of discrete tokens, and we train a neural network to perceive the image and generate the desired sequence. Our approach is based mainly on the intuition that if a neural network knows about where and what the objects are, we just need to teach it how to read them out. Beyond the use of task-specific data augmentations, our approach makes minimal assumptions about the task, yet it achieves competitive results on the challenging COCO dataset, compared to highly specialized and well optimized detection algorithms.

****

Geometric and Physical Quantities improve E(3) Equivariant Message Passing

Johannes Brandstetter,Rob Hesselink,Elise van der Pol,Erik J Bekkers,Max Welling

Including covariant information, such as position, force, velocity or spin is important in many tasks in computational physics and chemistry. We introduce Steerable E(3) Equivariant Graph Neural Networks (SEGNNs) that generalise equivariant graph networks, such that node and edge attributes are not restricted to invariant scalars, but can contain covariant information, such as vectors or tensors. Our model, composed of steerable MLPs, is able to incorporate geometric and physical information in both the message and update functions.

Through the definition of steerable node attributes, the MLPs provide a new class of activation functions for general use with steerable feature fields. We discuss ours and related work through the lens of equivariant non-linear convolutions, which further allows us to pin-point the successful components of SEGNNs: non-linear message aggregation improves upon classic linear (steerable) point convolutions; steerable messages improve upon recent equivariant graph networks that send invariant messages. We demonstrate the effectiveness of our method on several tasks in computational physics and chemistry and provide extensive ablation studies.

****

Particle Stochastic Dual Coordinate Ascent: Exponential convergent algorithm for mean field neural network optimization

Kazusato Oko,Taiji Suzuki,Atsushi Nitanda,Denny Wu

We introduce Particle-SDCA, a gradient-based optimization algorithm for two-layer neural networks in the mean field regime that achieves exponential convergence rate in regularized empirical risk minimization. The proposed algorithm can be regarded as an infinite dimensional extension of Stochastic Dual Coordinate Ascent (SDCA) in the probability space: we exploit the convexity of the dual problem, for which the coordinate-wise proximal gradient method can be applied. Our proposed method inherits advantages of the original SDCA, including (i) exponential convergence (with respect to the outer iteration steps), and (ii) better dependency on the sample size and condition number than the full-batch gradient method. One technical challenge in implementing the SDCA update is the intractable integral over the entire parameter space at every step. To overcome this limitation, we propose a tractable \textit{particle method} that approximately solves the dual problem, and an importance re-weighted technique to reduce the computational cost. The convergence rate of our method is verified by numerical experiments.

****

No Shifted Augmentations (NSA): strong baselines for self-supervised Anomaly Detection

Mohamed Yousef,Tom Bishop,Unmesh Kurup

Unsupervised Anomaly detection (AD) requires building a notion of normalcy, distinguishing in-distribution (ID) and out-of-distribution (OOD) data, using only available ID samples. Recently, large gains were made on this task for the domain of natural images using self-supervised contrastive feature learning as a first step followed by kNN or traditional one-class classifiers for feature scoring. Learned representations that are non-uniformly distributed on the unit hypersphere have been shown to be beneficial for this task. We go a step further and investigate how the \emph {geometrical compactness} of the ID feature distribution makes isolating and detecting outliers easier, especially in the realistic situation when ID training data is polluted (i.e. ID data contains some OOD data that is used for learning the feature extractor parameters).

We propose novel architectural modifications to the self-supervised feature learning step, that enable such compact ID distributions to be learned. We show that the proposed modifications can be effectively applied to most existing self-supervised learning objectives with large gains in performance. Furthermore, this improved OOD performance is obtained without resorting to tricks such as using strongly augmented ID images (e.g. by 90 degree rotations) as proxies for the unseen OOD data, which imposes overly prescriptive assumptions about ID data and its invariances.

We perform extensive studies on benchmark datasets for one-class OOD detection and show state-of-the-art performance in the presence of pollution in the ID data, and comparable performance otherwise. We also propose and extensively evaluate a novel feature scoring technique based on the angular Mahalanobis distance, and propose a simple and novel technique for feature ensembling during evaluation that enables a big boost in performance at nearly zero run-time cost compared to the standard use of model ensembling or test time augmentations. Code for all models and experiments will be made open-source.

****

Deep Learning of Intrinsically Motivated Options in the Arcade Learning Environment

Louis Bagot,Kevin Mets,Tom De Schepper,Peter Hellinckx,Steven Latre

Although Intrinsic Motivation allows a Reinforcement Learning agent to generate directed behaviors in an environment, even with sparse or noisy rewards, combining intrinsic and extrinsic rewards is non trivial. As an alternative to the widespread method of a weighted sum of rewards, Explore Options let the agent call an intrinsically motivated agent in order to observe and learn from interesting behaviors in the environment. Such options have only been established for simple tabular cases, and are unfit to high dimensional spaces. In this paper, we propose Deep Explore Options, revising Explore Options within the Deep Reinforcement Learning paradigm to tackle complex visual problems. Deep Explore Options can naturally learn from several unrelated intrinsic rewards, ignore harmful intrinsic rewards, learn to balance exploration, but also isolate exploitative or exploratory behaviors. In order to achieve this, we first introduce J-PER, a new transition-selection algorithm based on the interest of multiple agents. Next, we propose to consider intrinsic reward learning as an auxiliary task, with a resulting architecture achieving 50% faster wall-clock speed and building a stronger, shared representation. We test Deep Explore Options on hard and easy exploration games of the Atari Suite, following a benchmarking study to ensure fairness. Our results show that not only can they learn from multiple intrinsic rewards, they are a very strong alternative to a weighted sum of rewards, convincingly beating the baselines in 4 of the 6 tested environments, and with comparable performances in the other 2.

****

SphereFace2: Binary Classification is All You Need for Deep Face Recognition

Yandong Wen,Weiyang Liu,Adrian Weller,Bhiksha Raj,Rita Singh

State-of-the-art deep face recognition methods are mostly trained with a softmax-based multi-class classification framework. Despite being popular and effective, these methods still have a few shortcomings that limit empirical performance. In this paper, we start by identifying the discrepancy between training and evaluation in the existing multi-class classification framework and then discuss the potential limitations caused by the "competitive" nature of softmax normalization. Motivated by these limitations, we propose a novel binary classification training framework, termed SphereFace2. In contrast to existing methods, SphereFace2 circumvents the softmax normalization, as well as the corresponding closed-set assumption. This effectively bridges the gap between training and evaluation, enabling the representations to be improved individually by each binary classification task. Besides designing a specific well-performing loss function, we summarize a few general principles for this "one-vs-all" binary classification framework so that it can outperform current competitive methods. Our experiments on popular benchmarks demonstrate that SphereFace2 can consistently outperform state-of-the-art deep face recognition methods.

****

Counterbalancing Teacher: Regularizing Batch Normalized Models for Robustness

Saeid Asgari,Fereshte Khani,Ali Gholami,Kristy Choi,Linh Tran,Ran Zhang

Batch normalization (BN) is a ubiquitous technique for training deep neural networks that accelerates their convergence to reach higher accuracy. However, we demonstrate that BN comes with a fundamental drawback: it incentivizes the model to rely on frequent low-variance features that are highly specific to the training (in-domain) data, and thus fails to generalize to out-of-domain examples. In this work, we investigate this phenomenon by first showing that removing BN layers across a wide range of architectures leads to lower out-of-domain and corruption errors at the cost of higher in-domain error. We then propose the Counterbalancing Teacher (CT) method, which leverages a frozen copy of the same model without BN as a teacher to enforce the student network's learning of robust representations by substantially adapting its weights through a consistency loss function. This regularization signal helps CT perform well in unforeseen data shifts, even without information from the target domain as in prior works. We theoretically show in an overparameterized linear regression setting why normalization leads a model's reliance on such in-domain features, and empirically demonstrate the efficacy of CT by outperforming several methods on standard robustness benchmark datasets such as CIFAR-10-C, CIFAR-100-C, and VLCS.

****

The Effects of Invertibility on the Representational Complexity of Encoders in Variational Autoencoders

Divyansh Pareek,Andrej Risteski

Training and using modern neural-network based latent-variable generative models (like Variational Autoencoders) often require simultaneously training a generative direction along with an inferential (encoding) direction, which approximates the posterior distribution over the latent variables. Thus, the question arises: how complex does the inferential model need to be, in order to be able to accurately model the posterior distribution of a given generative model? In this paper, we identify an important property of the generative map impacting the required size of the encoder. We show that if the generative map is ``strongly invertible" (in a sense we suitably formalize), the inferential model need not be much more complex. Conversely, we prove that there exist non-invertible generative maps, for which the encoding direction needs to be exponentially larger (under standard assumptions in computational complexity). Importantly, we do not require the generative model to be layerwise invertible, which a lot of the related literature assumes and isn't satisfied by many architectures used in practice (e.g. convolution and pooling based networks). Thus, we provide theoretical support for the empirical wisdom that learning deep generative models is harder when data lies on a low-dimensional manifold.

****

Tracking the risk of a deployed model and detecting harmful distribution shifts

Aleksandr Podkopaev,Aaditya Ramdas

When deployed in the real world, machine learning models inevitably encounter changes in the data distribution, and certain---but not all---distribution shifts could result in significant performance degradation. In practice, it may make sense to ignore benign shifts, under which the performance of a deployed model does not degrade substantially, making interventions by a human expert (or model retraining) unnecessary. While several works have developed tests for distribution shifts, these typically either use non-sequential methods, or detect arbitrary shifts (benign or harmful), or both. We argue that a sensible method for firing off a warning has to both (a) detect harmful shifts while ignoring benign ones, and (b) allow continuous monitoring of model performance without increasing the false alarm rate. In this work, we design simple sequential tools for testing if the difference between source (training) and target (test) distributions leads to a significant increase in a risk function of interest, like accuracy or calibration. Recent advances in constructing time-uniform confidence sequences allow efficient aggregation of statistical evidence accumulated during the tracking process. The designed framework is applicable in settings where (some) true labels are revealed after the prediction is performed, or when batches of labels become available in a delayed fashion. We demonstrate the efficacy of the proposed framework through an extensive empirical study on a collection of simulated and real datasets.

****

A Modulation Layer to Increase Neural Network Robustness Against Data Quality Issues

Mohamed Abdelhack,Jiaming Zhang,Sandhya Tripathi,Bradley A Fritz,Michael Avidan,Yixin Chen,Christopher Ryan King

Data quality is a common problem in machine learning, especially in high-stakes settings such as healthcare. Missing data affects accuracy, calibration, and feature attribution in complex patterns. Developers often train models on carefully curated datasets to minimize missing data bias; however, this reduces the usability of such models in production environments, such as real-time healthcare records. Making machine learning models robust to missing data is therefore crucial for practical application. While some classifiers naturally handle missing data, others, such as deep neural networks, are not designed for unknown values. We propose a novel neural network modification to mitigate the impacts of missing data. The approach is inspired by neuromodulation that is performed by biological neural networks. Our proposal replaces the fixed weights of a fully-connected layer with a function of an additional input (reliability score) at each input, mimicking the ability of cortex to up- and down-weight inputs based on the presence of other data. The modulation function is jointly learned with the main task using a multi-layer perceptron. We tested our modulating fully connected layer on multiple classification, regression, and imputation problems, and it either improved performance or generated comparable performance to conventional neural network architectures concatenating reliability to the inputs. Models with modulating layers were more robust against degradation of data quality by introducing additional missingness at evaluation time. These results suggest that explicitly accounting for reduced information quality with a modulating fully connected layer can enable the deployment of artificial intelligence systems in real-time settings.

****

FedPAGE: A Fast Local Stochastic Gradient Method for Communication-Efficient Federated Learning

Haoyu Zhao,Zhize Li,Peter Richtárik

Federated Averaging (FedAvg, also known as Local-SGD) (McMahan et al., 2017) is a classical federated learning algorithm in which clients run multiple local SGD steps before communicating their update to an orchestrating server. We propose a new federated learning algorithm, FedPAGE, able to further reduce the communication complexity by utilizing the recent optimal PAGE method (Li et al., 2021) instead of plain SGD in FedAvg. We show that FedPAGE uses much fewer communicatio n rounds than previous local methods for both federated convex and nonconvex optimization. Concretely, 1) in the convex setting, the number of communication rounds of FedPAGE is $O(\frac{N^{3/4}}{S\epsilon})$, improving the best-known result $O(\frac{N}{S\epsilon})$ of SCAFFOLD (Karimireddy et al.,2020) by a factor of $N^{1/4}$, where N is the total number of clients (usually is very large in federated learning), S is the sampled subset of clients in each communication round, and ϵ is the target error; 2) in the nonconvex setting, the number of communication rounds of FedPAGE is $O(\frac{\sqrt{N}+S}{S\epsilon^2})$, improving the best-known result $O(\frac{N^{2/3}}{S^{2/3}\epsilon^2})$ of SCAFFOLD (Karimireddy et al.,2020) by a factor of $N^{1/6}S^{1/3}$, if the sampled clients $S\leq \sqrt{N}$. Note that in both settings, the communication cost for each round is the same for both FedPAGE and SCAFFOLD. As a result, FedPAGE achieves new state-of-the-art results in terms of communication complexity for both federated convex and nonconvex optimization.

****

Towards Understanding the Robustness Against Evasion Attack on Categorical Data

Hongyan Bao,Yufei Han,Yujun Zhou,Yun Shen,Xiangliang Zhang

Characterizing and assessing the adversarial vulnerability of classification models with categorical input has been a practically important, while rarely explored research problem. Our work echoes the challenge by first unveiling the impact factors of adversarial vulnerability of classification models with categorical data based on an information-theoretic adversarial risk analysis about the targeted classifier. Though certifying the robustness of such classification models is intrinsically an NP-hard combinatorial problem, our study shows that the robustness certification can be solved via an efficient greedy exploration of the discrete attack space for any measurable classifiers with a mild smoothness constraint. Our proposed robustness certification framework is instantiated with deep neural network models applied on real-world safety-critic data sources. Our empirical observations confirm the impact of the key adversarial risk factors with categorical input.

****

Learning Curves for SGD on Structured Features

Blake Bordelon,Cengiz Pehlevan

The generalization performance of a machine learning algorithm such as a neural network depends in a non-trivial way on the structure of the data distribution. To analyze the influence of data structure on test loss dynamics, we study an exactly solveable model of stochastic gradient descent (SGD) on the square loss which predicts test error when training on features with arbitrary covariance structure. We solve the theory exactly for both Gaussian features and arbitrary features and we show that the simpler Gaussian model accurately predicts test loss of nonlinear random-feature models and neural networks in the kernel regime trained with SGD on real datasets such as MNIST and CIFAR-10. We show that the optimal batch size at a fixed compute budget is typically small and depends on the feature correlation structure, demonstrating the computational benefits of SGD with small batch sizes. Lastly, we extend our theory to the more usual setting of stochastic gradient descent on a fixed subsampled training set, showing that both training and test error can be accurately predicted in our framework on real data.

****

NASViT: Neural Architecture Search for Efficient Vision Transformers with Gradient Conflict aware Supernet Training

Chengyue Gong,Dilin Wang,Meng Li,Xinlei Chen,Zhicheng Yan,Yuandong Tian,qiang liu,Vikas Chandra

Designing accurate and efficient vision transformers (ViTs) is a highly important but challenging task. Supernet-based one-shot neural architecture search (NAS) enables fast architecture optimization and has achieved state-of-the-art (SOTA) results on convolutional neural networks (CNNs). However, directly applying the supernet-based NAS to optimize ViTs leads to poor performance - even worse compared to training single ViTs. In this work, we observe that the poor performance is due to a gradient conflict issue: the gradients of different sub-networks co

nflict with that of the supernet more severely in ViTs than CNNs, which leads to early saturation in training and inferior convergence. To alleviate this issue, we propose a series of techniques, including a gradient projection algorithm, a switchable layer scaling design, and a simplified data augmentation and regularization training recipe. The proposed techniques significantly improve the convergence and the performance of all sub-networks. Our discovered hybrid ViT model family, dubbed NASViT, achieves top-1 accuracy from 78.2% to 81.8% on ImageNet from 200M to 800M FLOPs, and outperforms all the prior art CNNs and ViTs, including AlphaNet and LeViT, etc. When transferred to semantic segmentation tasks, NASViTs also outperform previous backbones on both Cityscape and ADE20K datasets, achieving 73.2% and 37.9% mIoU with only 5G FLOPs, respectively. Code is available at
https://github.com/facebookresearch/NASViT.

****

Graphon based Clustering and Testing of Networks: Algorithms and Theory

Mahalakshmi Sabanayagam,Leena Chennuru Vankadara,Debarghya Ghoshdastidar

Network-valued data are encountered in a wide range of applications, and pose challenges in learning due to their complex structure and absence of vertex correspondence. Typical examples of such problems include classification or grouping of protein structures and social networks. Various methods, ranging from graph kernels to graph neural networks, have been proposed that achieve some success in graph classification problems. However, most methods have limited theoretical justification, and their applicability beyond classification remains unexplored. In this work, we propose methods for clustering multiple graphs, without vertex correspondence, that are inspired by the recent literature on estimating graphons---symmetric functions corresponding to infinite vertex limit of graphs. We propose a novel graph distance based on sorting-and-smoothing graphon estimators. Using the proposed graph distance, we present two clustering algorithms and show that they achieve state-of-the-art results. We prove the statistical consistency of both algorithms under Lipschitz assumptions on the graph degrees. We further study the applicability of the proposed distance for graph two-sample testing problems.

****

Escaping Stochastic Traps with Aleatoric Mapping Agents

Augustine N. Mavor-Parker,Kimberly A Young,Caswell Barry,Lewis Griffin

When extrinsic rewards are sparse, artificial agents struggle to explore an environment. Curiosity, implemented as an intrinsic reward for prediction errors, can improve exploration but fails when faced with action-dependent noise sources. We present aleatoric mapping agents (AMAs), a neuroscience inspired solution modeled on the cholinergic system of the mammalian brain. AMAs aim to explicitly ascertain which dynamics of the environment are unpredictable, regardless of whether those dynamics are induced by the actions of the agent. This is achieved by generating separate forward predictions for the mean and aleatoric uncertainty of future states with reducing intrinsic rewards for those states that are unpredictable. We show AMAs are able to effectively circumvent action-dependent stochastic traps that immobilise conventional curiosity driven agents.

****

Network Augmentation for Tiny Deep Learning

Han Cai,Chuang Gan,Ji Lin,song han

We introduce Network Augmentation (NetAug), a new training method for improving the performance of tiny neural networks. Existing regularization techniques (e.g., data augmentation, dropout) have shown much success on large neural networks by adding noise to overcome over-fitting. However, we found these techniques hurt the performance of tiny neural networks. We argue that training tiny models are different from large models: rather than augmenting the data, we should augment the model, since tiny models tend to suffer from under-fitting rather than over-fitting due to limited capacity. To alleviate this issue, NetAug augments the network (reverse dropout) instead of inserting noise into the dataset or the network. It puts the tiny model into larger models and encourages it to work as a s

ub-model of larger models to get extra supervision, in addition to functioning as an independent model. At test time, only the tiny model is used for inference, incurring zero inference overhead. We demonstrate the effectiveness of NetAug on image classification and object detection. NetAug consistently improves the performance of tiny models, achieving up to 2.2% accuracy improvement on ImageNet. On object detection, achieving the same level of performance, NetAug requires 41% fewer MACs on Pascal VOC and 38% fewer MACs on COCO than the baseline.
****

Bridging the Gap: Providing Post-Hoc Symbolic Explanations for Sequential Decision-Making Problems with Inscrutable Representations
Sarath Sreedharan,Utkarsh Soni,Mudit Verma,Siddharth Srivastava,Subbarao Kambhampati
As increasingly complex AI systems are introduced into our daily lives, it becomes important for such systems to be capable of explaining the rationale for their decisions and allowing users to contest these decisions. A significant hurdle to allowing for such explanatory dialogue could be the {\em vocabulary mismatch} between the user and the AI system. This paper introduces methods for providing contrastive explanations in terms of user-specified concepts for sequential decision-making settings where the system's model of the task may be best represented as an inscrutable model. We do this by building partial symbolic models of a local approximation of the task that can be leveraged to answer the user queries. We test these methods on a popular Atari game (Montezuma's Revenge) and variants of Sokoban (a well-known planning benchmark) and report the results of user studies to evaluate whether people find explanations generated in this form useful.
****

Distributional Reinforcement Learning with Monotonic Splines
Yudong Luo,Guiliang Liu,Haonan Duan,Oliver Schulte,Pascal Poupart
Distributional Reinforcement Learning (RL) differs from traditional RL by estimating the distribution over returns to capture the intrinsic uncertainty of MDPs. One key challenge in distributional RL lies in how to parameterize the quantile function when minimizing the Wasserstein metric of temporal differences. Existing algorithms use step functions or piecewise linear functions. In this paper, we propose to learn smooth continuous quantile functions represented by monotonic rational-quadratic splines, which also naturally solve the quantile crossing problem. Experiments in stochastic environments show that a dense estimation for quantile functions enhances distributional RL in terms of faster empirical convergence and higher rewards in most cases.
****

MQTransformer: Multi-Horizon Forecasts with Context Dependent and Feedback-Aware Attention
Carson Eisenach,Yagna Patel,Dhruv Madeka
Recent advances in neural forecasting have produced major improvements in accuracy for probabilistic demand prediction. In this work, we propose novel improvements to the current state of the art by incorporating changes inspired by recent advances in Transformer architectures for Natural Language Processing. We develop a novel decoder-encoder attention for context-alignment, improving forecasting accuracy by allowing the network to study its own history based on the context for which it is producing a forecast. We also present a novel positional encoding that allows the neural network to learn context-dependent seasonality functions as well as arbitrary holiday distances. Finally we show that the current state of the art MQ-Forecaster (Wen et al., 2017) models display excess variability by failing to leverage previous errors in the forecast to improve accuracy. We propose a novel decoder-self attention scheme for forecasting that produces significant improvements in the excess variation of the forecast.
****

RAR: Region-Aware Point Cloud Registration
Yu Hao,Yi Fang
This paper concerns the research problem of point cloud registration to find the rigid transformation to optimally align the source point set with the target on

e. Learning robust point cloud registration models with deep neural networks has emerged as a powerful paradigm, offering promising performance in predicting the global geometric transformation for a pair of point sets. Existing methods firstly leverage an encoder to regress a latent shape embedding, which is then decoded into a shape-conditioned transformation via concatenation-based conditioning. However, different regions of a 3D shape vary in their geometric structures which makes it more sense that we have a region-conditioned transformation instead of the shape-conditioned one. With this observation, in this paper we present a \underline{R}egion-\underline{A}ware point cloud \underline{R}egistration, denoted as RAR, to predict transformation for pairwise point sets in the self-supervised learning fashion. More specifically, we develop a novel region-aware decoder (RAD) module that is formed with an implicit neural region representation parameterized by neural networks. The implicit neural region representation is learned with a self-supervised 3D shape reconstruction loss without the need for region labels. Consequently, the region-aware decoder (RAD) module guides the training of the region-aware transformation (RAT) module and region-aware weight (RAW) module, which predict the transforms and weights for different regions respectively. The global geometric transformation from source point set to target one is then formed by the weighted fusion of region-aware transforms. Compared to the state-of-the-art approaches, our experiments show that our RAR achieves superior registration performance over various benchmark datasets (e.g. ModelNet40).

****

Occupy & Specify: Investigations into a Maximum Credit Assignment Occupancy Objective for Data-efficient Reinforcement Learning

Emmanuel Daucé

The capability to widely sample the state and action spaces is a key ingredient toward building effective reinforcement learning algorithms. The trade-off between exploration and exploitation generally requires the use of a data model, from which novelty bonuses are estimated and used to bias the return toward wider exploration. Surprisingly, little is known about the optimization objective followed when novelty (or entropy) bonuses are considered. Following the ``probability matching'' principle, we interpret here returns (cumulative rewards) as set points that fixate the occupancy of the state space, that is the frequency at which the different states are expected to be visited during trials. The circular dependence of the rewards sampling on the occupancy/policy makes it difficult to evaluate. We provide here a variational formulation for the matching objective, named MaCAO (Maximal Credit Assignment Occupancy) that interprets rewards as a log-likelihood on occupancy, that operates anticausally from the effects toward the causes. It is, broadly speaking, an estimation of the contribution of a state toward reaching a (future) goal. It is constructed so as to provide better convergence guaranties, with a complementary term serving as a regularizer, that, in principle, may reduce the greediness. In the absence of an explicit target occupancy, a uniform prior is used, making the regularizer consistent with a MaxEnt (Maximum Entropy) objective on states. Optimizing the entropy on states in known to be more tricky than optimizing the entropy on actions, because of an external sampling through the (unknown) environment, that prevents the propagation of a gradient. In our practical implementations, the MaxEnt regularizer is interpreted as a TD-error rather than a reward, making it possible to define an update in both the discrete and continuous cases. It is implemented on an actor-critic off-policy setup with a replay buffer, using gradient descent on a multi-layered neural network, and shown to provide significant increase in the sampling efficacy, that reflects in a reduced training time and higher returns on a set of classical motor learning benchmarks, in both the dense and the sparse rewards cases.

****

Two Birds, One Stone: Achieving both Differential Privacy and Certified Robustness for Pre-trained Classifiers via Input Perturbation

Pengfei Tang,Wenjie Wang,Xiaolan Gu,Jian Lou,Li Xiong,Ming Li

Recent studies have shown that pre-trained classifiers are increasingly powerful to improve the performance on different tasks, e.g, neural language processing, image classification. However, adversarial examples from attackers can trick pr e-trained classifiers to misclassify. To solve this challenge, a reconstruction network is built before the public pre-trained classifiers to offer certified robustness and defend against adversarial examples through input perturbation. On the other hand, the reconstruction network requires training on the dataset, which incurs privacy leakage of training data through inference attacks. To prevent this leakage, differential privacy (DP) is applied to offer a provable privacy guarantee on training data through gradient perturbation. Most existing works employ certified robustness and DP independently and fail to exploit the fact that input perturbation designed to achieve certified robustness can achieve (partial) DP. In this paper, we propose perturbation transformation to show how the input perturbation designed for certified robustness can be transformed into gradient perturbation during training. We propose Multivariate Gaussian mechanism to analyze the privacy guarantee of this transformed gradient perturbation and precisely quantify the level of DP achieved by input perturbation. To satisfy the overall DP requirement, we add additional gradient perturbation during training and propose Mixed Multivariate Gaussian Analysis to analyze the privacy guarantee provided by the transformed gradient perturbation and additional gradient perturbation. Moreover, we prove that Mixed Multivariate Gaussian Analysis can work with moments accountant to provide a tight DP estimation. Extensive experiments on benchmark datasets show that our framework significantly outperforms state-of-the-art methods and achieves better accuracy and robustness under the same privacy guarantee.

****

Dual Training of Energy-Based Models with Overparametrized Shallow Neural Networks

Carles Domingo-Enrich,Alberto Bietti,Marylou Gabrié,Joan Bruna,Eric Vanden-Eijnden

Energy-based models (EBMs) are generative models that are usually trained via maximum likelihood estimation. This approach becomes challenging in generic situations where the trained energy is nonconvex, due to the need to sample the Gibbs distribution associated with this energy. Using general Fenchel duality results, we derive variational principles dual to maximum likelihood EBMs with shallow overparametrized neural network energies, both in the active (aka feature-learning) and lazy regimes. In the active regime, this dual formulation leads to a training algorithm in which one updates concurrently the particles in the sample space and the neurons in the parameter space of the energy at a faster rate. We also consider a variant of this algorithm in which the particles are sometimes restarted at random samples drawn from the data set, and show that performing these restarts at every iteration step corresponds to score matching training. Using intermediate parameter setups in our dual algorithm thereby gives a way to interpolate between maximum likelihood and score matching training. These results are illustrated in simple numerical experiments.

****

Gotta Go Fast When Generating Data with Score-Based Models

Alexia Jolicoeur-Martineau,Ke Li,Rémi Piché-Taillefer,Tal Kachman,Ioannis Mitliagkas

Score-based (denoising diffusion) generative models have recently gained a lot of success in generating realistic and diverse data. These approaches define a forward diffusion process for transforming data to noise and generate data by reversing it (thereby going from noise to data). Unfortunately, current score-based models generate data very slowly due to the sheer number of score network evaluations required by numerical SDE solvers.

In this work, we aim to accelerate this process by devising a more efficient SDE solver. Existing approaches rely on the Euler-Maruyama (EM) solver, which uses a fixed step size. We found that naively replacing it with other SDE solvers fares poorly - they either result in low-quality samples or become slower than EM. To get around this issue, we carefully devise an SDE solver with adaptive step sizes tailored to score-based generative models piece by piece. Our solver requires only two score function evaluations, rarely rejects samples, and leads to hig h-quality samples. Our approach generates data 2 to 10 times faster than EM while achieving better or equal sample quality. For high-resolution images, our method leads to significantly higher quality samples than all other methods tested. Our SDE solver has the benefit of requiring no step size tuning.

****

Learning Graph Structure from Convolutional Mixtures

Max Wasserman,Saurabh Sihag,Gonzalo Mateos,Alejandro Ribeiro

Machine learning frameworks such as graph neural networks typically rely on a given, fixed graph to exploit relational inductive biases and thus effectively learn from network data. However, assuming the knowledge of said graphs may be untenable in practice, which motivates the problem of inferring graph structure from data. In this paper, we postulate a graph convolutional relationship between the observed and latent graphs, and formulate the graph learning task as a network inverse (deconvolution) problem. In lieu of eigendecomposition-based spectral methods or iterative optimization solutions, we unroll and truncate proximal gradient iterations to arrive at a parameterized neural network architecture that we call a Graph Deconvolution Network (GDN). GDNs can learn a distribution of graphs in a supervised fashion, and perform link-prediction or edge-weight regression tasks by adapting the loss function. Since layers directly operate on, combine, and refine graph objects (instead of node features), GDNs are inherently inductive and can generalize to larger-sized graphs after training.

Algorithm unrolling offers an explicit handle on computational complexity; we trade-off training time in return for quick approximations to the inverse problem solution, obtained via a forward pass through the learnt model. We corroborate GDN's superior graph recovery performance using synthetic data in supervised settings, as well as its ability to generalize to graphs orders of magnitude larger that those seen in training. Using the Human Connectome Project-Young Adult neuroimaging dataset, we demonstrate the robustness and representation power of our model by inferring structural brain networks from functional connectivity estimated using fMRI signals.

****

Toward Faithful Case-based Reasoning through Learning Prototypes in a Nearest Neighbor-friendly Space.

Seyed Omid Davoudi,Majid Komeili

Recent advances in machine learning have brought opportunities for the ever-increasing use of AI in the real world. This has created concerns about the black-box nature of many of the most recent machine learning approaches. In this work, we propose an interpretable neural network that leverages metric and prototype learning for classification tasks. It encodes its own explanations and provides an improved case-based reasoning through learning prototypes in an embedding space learned by a probabilistic nearest neighbor rule. Through experiments, we demonstrated the effectiveness of the proposed method in both performance and the accuracy of the explanations provided.

****

Augmented Sliced Wasserstein Distances

Xiongjie Chen,Yongxin Yang,Yunpeng Li

While theoretically appealing, the application of the Wasserstein distance to large-scale machine learning problems has been hampered by its prohibitive computational cost. The sliced Wasserstein distance and its variants improve the computational efficiency through the random projection, yet they suffer from low accuracy if the number of projections is not sufficiently large, because the majority of projections result in trivially small values. In this work, we propose a new family of distance metrics, called augmented sliced Wasserstein distances (ASWDs), constructed by first mapping samples to higher-dimensional hypersurfaces parameterized by neural networks. It is derived from a key observation that (random) linear projections of samples residing on these hypersurfaces would translate to much more flexible nonlinear projections in the original sample space, so they can capture complex structures of the data distribution. We show that the hypersurfaces can be optimized by gradient ascent efficiently. We provide the condition under which the ASWD is a valid metric and show that this can be obtained by an injective neural network architecture. Numerical results demonstrate that the ASWD significantly outperforms other Wasserstein variants for both synthetic and real-world problems.

****

Relational Learning with Variational Bayes

Kuang-Hung Liu

In psychology, relational learning refers to the ability to recognize and respond to relationship among objects irrespective of the nature of those objects. Relational learning has long been recognized as a hallmark of human cognition and a key question in artificial intelligence research. In this work, we propose an unsupervised learning method for addressing the relational learning problem where we learn the underlying relationship between a pair of data irrespective of the nature of those data. The central idea of the proposed method is to encapsulate the relational learning problem with a probabilistic graphical model in which we perform inference to learn about data relationship and other relational processing tasks.

****

Boosting Randomized Smoothing with Variance Reduced Classifiers

Miklós Z. Horváth,Mark Niklas Mueller,Marc Fischer,Martin Vechev

Randomized Smoothing (RS) is a promising method for obtaining robustness certificates by evaluating a base model under noise. In this work, we: (i) theoretically motivate why ensembles are a particularly suitable choice as base models for RS, and (ii) empirically confirm this choice, obtaining state-of-the-art results in multiple settings. The key insight of our work is that the reduced variance of ensembles over the perturbations introduced in RS leads to significantly more consistent classifications for a given input. This, in turn, leads to substantially increased certifiable radii for samples close to the decision boundary. Additionally, we introduce key optimizations which enable an up to 55-fold decrease in sample complexity of RS for predetermined radii, thus drastically reducing its computational overhead. Experimentally, we show that ensembles of only 3 to 10 classifiers consistently improve on their strongest constituting model with respect to their average certified radius (ACR) by 5% to 21% on both CIFAR10 and ImageNet, achieving a new state-of-the-art ACR of 0.86 and 1.11, respectively. We release all code and models required to reproduce our results at https://github.com/eth-sri/smoothing-ensembles.

****

InstaHide’s Sample Complexity When Mixing Two Private Images

Baihe Huang,Zhao Song,Runzhou Tao,Ruizhe Zhang,Danyang Zhuo

Inspired by InstaHide challenge [Huang, Song, Li and Arora'20], [Chen, Song and Zhuo'20] recently provides one mathematical formulation of InstaHide attack problem under Gaussian images distribution. They show that it suffices to use $O(n_{\mathsf{priv}}^{k_{\mathsf{priv}} - 2/(k_{\mathsf{priv}} + 1)})$ samples to recover one private image in $n_{\mathsf{priv}}^{O(k_{\mathsf{priv}})} + \mathrm{poly}(n_{\mathsf{pub}})$ time for any integer k_{priv}, where n_{priv} and n_{pub} denote the number of images used in the private and the public dataset to generate a mixed image sample. Under the current setup for the InstaHide challenge of mixing two private images ($k_{\mathsf{priv}} = 2$), this means $n_{\mathsf{priv}}^{4/3}$ samples are sufficient to recover a private image. In this work, we show that $n_{\mathsf{priv}} \log (n_{\mathsf{priv}})$ samples are sufficient (information-theoretically) for recovering all the private images.

****

Provably Robust Adversarial Examples

Dimitar Iliev Dimitrov,Gagandeep Singh,Timon Gehr,Martin Vechev

We introduce the concept of provably robust adversarial examples for deep neural networks – connected input regions constructed from standard adversarial examples which are guaranteed to be robust to a set of real-world perturbations (such as changes in pixel intensity and geometric transformations). We present a novel method called PARADE for generating these regions in a scalable manner which works by iteratively refining the region initially obtained via sampling until a refined region is certified to be adversarial with existing state-of-the-art verifiers. At each step, a novel optimization procedure is applied to maximize the region's volume under the constraint that the convex relaxation of the network behavior with respect to the region implies a chosen bound on the certification objective. Our experimental evaluation shows the effectiveness of PARADE: it successfully finds large provably robust regions including ones containing $\approx 10^{573}$ adversarial examples for pixel intensity and $\approx 10^{599}$ for geometric perturbations. The provability enables our robust examples to be significantly more effective against state-of-the-art defenses based on randomized smoothing than the individual attacks used to construct the regions.

****

Unsupervised Learning of Neurosymbolic Encoders

Eric Zhan,Jennifer J. Sun,Ann Kennedy,Yisong Yue,Swarat Chaudhuri

We present a framework for the unsupervised learning of neurosymbolic encoders, i.e., encoders obtained by composing neural networks with symbolic programs from a domain-specific language. Such a framework can naturally incorporate symbolic expert knowledge into the learning process and lead to more interpretable and factorized latent representations than fully neural encoders. Also, models learned this way can have downstream impact, as many analysis workflows can benefit from having clean programmatic descriptions. We ground our learning algorithm in the variational autoencoding (VAE) framework, where we aim to learn a neurosymbolic encoder in conjunction with a standard decoder. Our algorithm integrates standard VAE-style training with modern program synthesis techniques. We evaluate our method on learning latent representations for real-world trajectory data from animal biology and sports analytics. We show that our approach offers significantly better separation than standard VAEs and leads to practical gains on downstream tasks.

****

Joint Shapley values: a measure of joint feature importance

Chris Harris,Richard Pymar,Colin Rowat

The Shapley value is one of the most widely used measures of feature importance partly as it measures a feature's average effect on a model's prediction. We introduce joint Shapley values, which directly extend Shapley's axioms and intuitions: joint Shapley values measure a set of features' average effect on a model's prediction. We prove the uniqueness of joint Shapley values, for any order of explanation. Results for games show that joint Shapley values present different insights from existing interaction indices, which assess the effect of a feature within a set of features. The joint Shapley values seem to provide sensible results in ML attribution problems. With binary features, we present a presence-adjusted global value that is more consistent with local intuitions than the usual approach.

****

Low-Budget Active Learning via Wasserstein Distance: An Integer Programming Approach

Rafid Mahmood,Sanja Fidler,Marc T Law

Active learning is the process of training a model with limited labeled data by selecting a core subset of an unlabeled data pool to label. The large scale of data sets used in deep learning forces most sample selection strategies to employ efficient heuristics. This paper introduces an integer optimization problem for selecting a core set that minimizes the discrete Wasserstein distance from the unlabeled pool. We demonstrate that this problem can be tractably solved with a Generalized Benders Decomposition algorithm. Our strategy uses high-quality latent features that can be obtained by unsupervised learning on the unlabeled pool. Numerical results on several data sets show that our optimization approach is competitive with baselines and particularly outperforms them in the low budget regime where less than one percent of the data set is labeled.

****

Efficient Self-supervised Vision Transformers for Representation Learning

Chunyuan Li,Jianwei Yang,Pengchuan Zhang,Mei Gao,Bin Xiao,Xiyang Dai,Lu Yuan,Jianfeng Gao

This paper investigates two techniques for developing efficient self-supervised vision transformers (EsViT) for visual representation learning. First, we show through a comprehensive empirical study that multi-stage architectures with sparse self-attentions can significantly reduce modeling complexity but with a cost of losing the ability to capture fine-grained correspondences between image regions. Second, we propose a new pre-training task, non-contrastive region-matching, which allows the model to capture fine-grained region dependencies and as a result significantly improves the quality of the learned vision representations. Our results show that combining the two techniques, EsViT achieves 81.3% top-1 on the ImageNet linear probe evaluation, outperforming prior arts with around an order magnitude of higher throughput. When transferring to downstream linear classification tasks, EsViT outperforms its supervised counterpart on 17 out of 18 datasets. The code and pre-trained models are released at: https://github.com/microsoft/esvit

****

Exact Stochastic Newton Method for Deep Learning: the feedforward networks case.

Fares B. Mehouachi,Chaouki Kasmi

The inclusion of second-order information into Deep Learning optimization has drawn consistent interest as a way forward to improve upon gradient descent methods. Estimating the second-order update is often convoluted and computationally expensive, which drastically limits its usage scope and forces the use of various truncations and approximations.

This work demonstrates that it is possible to solve the Newton direction in the stochastic case exactly. We consider feedforward networks as a base model, build a second-order Lagrangian which we call Sifrian, and provide a closed-form formula for the exact stochastic Newton direction under some monotonicity and regularization conditions. We propose a convexity correction to escape saddle points, and we reconsider the intrinsic stochasticity of the online learning process to improve upon the formulas. We finally compare the performance of the developed solution with well-established training methods and show its viability as a training method for Deep Learning.

****

Visual Representation Learning Does Not Generalize Strongly Within the Same Domain

Lukas Schott,Julius Von Kügelgen,Frederik Träuble,Peter Vincent Gehler,Chris Russell,Matthias Bethge,Bernhard Schölkopf,Francesco Locatello,Wieland Brendel

An important component for generalization in machine learning is to uncover underlying latent factors of variation as well as the mechanism through which each factor acts in the world.

In this paper, we test whether 17 unsupervised, weakly supervised, and fully supervised representation learning approaches correctly infer the generative factors of variation in simple datasets (dSprites, Shapes3D, MPI3D) from controlled environments, and on our contributed CelebGlow dataset.

In contrast to prior robustness work that introduces novel factors of variation during test time, such as blur or other (un)structured noise, we here recompose, interpolate, or extrapolate only existing factors of variation from the training data set (e.g., small and medium-sized objects during training and large objects during testing). Models that learn the correct mechanism should be able to generalize to this benchmark.

In total, we train and test 2000+ models and observe that all of them struggle to learn the underlying mechanism regardless of supervision signal and architectural bias. Moreover, the generalization capabilities of all tested models drop significantly as we move from artificial datasets towards more realistic real-world datasets.

Despite their inability to identify the correct mechanism, the models are quite modular as their ability to infer other in-distribution factors remains fairly stable, providing only a single factor is out-of-distribution. These results point to an important yet understudied problem of learning mechanistic models of obs ervations that can facilitate generalization.

****

Learning-to-Count by Learning-to-Rank: Weakly Supervised Object Counting & Localization Using Only Pairwise Image Rankings

Adriano C. D'Alessandro,Ali Mahdavi Amiri,Ghassan Hamarneh

Object counting and localization in dense scenes is a challenging class of image analysis problems that typically requires labour intensive annotations to learn to solve. We propose a form of weak supervision that only requires object-based pairwise image rankings. These annotations can be collected rapidly with a single click per image pair and supply a weak signal for object quantity. However, the problem of actually extracting object counts and locations from rankings is challenging. Thus, we introduce adversarial density map generation, a strategy for regularizing the features of a ranking network such that the features correspond to an object proposal map where each proposal must be a Gaussian blob that integrates to 1. This places a soft integer and soft localization constraint on the representation, which encourages the network to satisfy the provided ranking constraints by detecting objects. We then demonstrate the effectiveness of our method for exploiting pairwise image rankings as a weakly supervised signal for object counting and localization on several datasets, and show results with a performance that approaches that of fully supervised methods on many counting benchmark datasets while relying on data that can be collected with a fraction of the annotation burden.

****

Neural Collapse Under MSE Loss: Proximity to and Dynamics on the Central Path

X.Y. Han,Vardan Papyan,David L. Donoho

The recently discovered Neural Collapse (NC) phenomenon occurs pervasively in today's deep net training paradigm of driving cross-entropy (CE) loss towards zero. During NC, last-layer features collapse to their class-means, both classifiers and class-means collapse to the same Simplex Equiangular Tight Frame, and classifier behavior collapses to the nearest-class-mean decision rule. Recent works demonstrated that deep nets trained with mean squared error (MSE) loss perform comparably to those trained with CE. As a preliminary, we empirically establish that NC emerges in such MSE-trained deep nets as well through experiments on three canonical networks and five benchmark datasets. We provide, in a Google Colab notebook, PyTorch code for reproducing MSE-NC and CE-NC: https://colab.research.google.com/github/neuralcollapse/neuralcollapse/blob/main/neuralcollapse.ipynb. The analytically-tractable MSE loss offers more mathematical opportunities than the hard-to-analyze CE loss, inspiring us to leverage MSE loss towards the theoretical investigation of NC. We develop three main contributions: (I) We show a new decomposition of the MSE loss into (A) terms directly interpretable through the lens of NC and which assume the last-layer classifier is exactly the least-squares classifier; and (B) a term capturing the deviation from this least-squares classifier. (II) We exhibit experiments on canonical datasets and networks demonstrating that term-(B) is negligible during training. This motivates us to introduce a new theoretical construct: the central path, where the linear classifier stays MSE-optimal for feature activations throughout the dynamics. (III) By studying renormalized gradient flow along the central path, we derive exact dynamics that predict NC.

****

Short-term memory in neural language models

Kristijan Armeni,Christopher Honey,Tal Linzen

When a language model is trained to predict natural language sequences, its prediction at each moment depends on a representation of prior context. Thus, language models require mechanisms to maintain and access memory. Although we design the architectural features of these models, we do not know how their memory systems are functionally organized via learning: what kind of information about the prior context can they retrieve? We reasoned that access to arbitrary individual tokens from the past could be computationally powerful, akin to the working memory which is important for flexible cognition in humans, and we therefore tested whether language models could ``retrieve'' the exact words that occurred previo usly in a text. In particular, we tested how the ability to retrieve prior words depended on (i) the number of words being retrieved, (ii) their semantic coherence, and (iii) the length and quality of the intervening text. We evaluated two particular architectures of neural language models: the attention-based transformer and the long short-term memory network (LSTM). In our paradigm, language models processed English text in which a list of nouns occurred twice. We operationalized retrieval as the reduction in surprisal from the first presentation of the list to its second presentation. We found that the transformer models retrieved both the identity and ordering of nouns from the first list. The transformer was successful even when the noun lists were semantically incoherent, and this effect was largely robust to the type or length of the intervening text. Further, the transformer’s retrieval was markedly enhanced when it was trained on a larger corpus and with greater model depth. Lastly, its ability to index prior tokens was dependent on learned attention patterns. In contrast, the LSTM models exhibited less precise retrieval (smaller reductions in surprisal). The LSTM’s retrieval was limited to list-initial tokens, and occurred only across short intervening texts. Moreover, the LSTM's retrieval was not sensitive to the order of nouns and this non-specific retrieval improved when the list was semantically coherent. In sum, the transformer, when trained to predict linguistic tokens, implements something akin to a working memory system, as it could flexibly retrieve individual token representations across arbitrary delays. Conversely, the LSTM maintained a coarser and more rapidly-decaying semantic gist of prior tokens, weighted heavily toward the earliest items. Thus, although the transformer and LSTM architectures were both trained to predict language sequences, only the transformer learned to flexibly index prior tokens.

****

Hidden Convexity of Wasserstein GANs: Interpretable Generative Models with Closed-Form Solutions

Arda Sahiner,Tolga Ergen,Batu Ozturkler,Burak Bartan,John M. Pauly,Morteza Mardani,Mert Pilanci

Generative Adversarial Networks (GANs) are commonly used for modeling complex distributions of data. Both the generators and discriminators of GANs are often modeled by neural networks, posing a non-transparent optimization problem which is non-convex and non-concave over the generator and discriminator, respectively. Such networks are often heuristically optimized with gradient descent-ascent (GDA), but it is unclear whether the optimization problem contains any saddle points, or whether heuristic methods can find them in practice. In this work, we analyze the training of Wasserstein GANs with two-layer neural network discriminators through the lens of convex duality, and for a variety of generators expose the conditions under which Wasserstein GANs can be solved exactly with convex optimization approaches, or can be represented as convex-concave games. Using this convex duality interpretation, we further demonstrate the impact of different activation functions of the discriminator. Our observations are verified with numerical results demonstrating the power of the convex interpretation, with an application in progressive training of convex architectures corresponding to linear generators and quadratic-activation discriminators for CelebA image generation. The code for our experiments is available at https://github.com/ardasahiner/ProCoGAN.

****

Memory Augmented Optimizers for Deep Learning

Paul-Aymeric Martin McRae,Prasanna Parthasarathi,Mido Assran,Sarath Chandar

Popular approaches for minimizing loss in data-driven learning often involve an abstraction or an explicit retention of the history of gradients for efficient parameter updates.

The aggregated history of gradients nudges the parameter updates in the right direction even when the gradients at any given step are not informative.

Although the history of gradients summarized in meta-parameters or explicitly stored in memory has been shown effective in theory and practice, the question of whether all or only a subset of the gradients in the history are sufficient in deciding the parameter updates remains unanswered.

In this paper, we propose a framework of memory-augmented gradient descent optimizers that retain a limited view of their gradient history in their internal memory.
Such optimizers scale well to large real-life datasets, and our experiments show that the memory augmented extensions of standard optimizers enjoy accelerated convergence and improved performance on a majority of computer vision and language tasks that we considered.
Additionally, we prove that the proposed class of optimizers with fixed-size memory converge under assumptions of strong convexity, regardless of which gradients are selected or how they are linearly combined to form the update step.
****

Weighted Training for Cross-Task Learning
Shuxiao Chen,Koby Crammer,Hangfeng He,Dan Roth,Weijie J Su
In this paper, we introduce Target-Aware Weighted Training (TAWT), a weighted training algorithm for cross-task learning based on minimizing a representation-based task distance between the source and target tasks. We show that TAWT is easy to implement, is computationally efficient, requires little hyperparameter tuning, and enjoys non-asymptotic learning-theoretic guarantees. The effectiveness of TAWT is corroborated through extensive experiments with BERT on four sequence tagging tasks in natural language processing (NLP), including part-of-speech (PoS) tagging, chunking, predicate detection, and named entity recognition (NER). As a byproduct, the proposed representation-based task distance allows one to reason in a theoretically principled way about several critical aspects of cross-task learning, such as the choice of the source data and the impact of fine-tuning.
****

Momentum Contrastive Autoencoder: Using Contrastive Learning for Latent Space Distribution Matching in WAE
Devansh Arpit,Aadyot Bhatnagar,Huan Wang,Caiming Xiong
Wasserstein autoencoder (WAE) shows that matching two distributions is equivalent to minimizing a simple autoencoder (AE) loss under the constraint that the latent space of this AE matches a pre-specified prior distribution. This latent space distribution matching is a core component of WAE, and a challenging task. In this paper, we propose to use the contrastive learning framework that has been shown to be effective for self-supervised representation learning, as a means to resolve this problem. We do so by exploiting the fact that contrastive learning objectives optimize the latent space distribution to be uniform over the unit hyper-sphere, which can be easily sampled from.
We show that using the contrastive learning framework to optimize the WAE loss achieves faster convergence and more stable optimization compared with existing popular algorithms for WAE. This is also reflected in the FID scores on CelebA and CIFAR-10 datasets, and the realistic generated image quality on the CelebA-HQ dataset.
****

Retrieval-Augmented Reinforcement Learning
Anirudh Goyal,Abram L. Friesen,Theophane Weber,Andrea Banino,Nan Rosemary Ke,Adria Puigdomenech Badia,Ksenia Konyushkova,Michal Valko,Simon Osindero,Timothy P Lillicrap,Nicolas Heess,Charles Blundell
Most deep reinforcement learning (RL) algorithms distill experience into parametric behavior policies or value functions via gradient updates. While effective, this approach has several disadvantages: (1) it is computationally expensive, (2) it can take many updates to integrate experiences into the parametric model, (3) experiences that are not fully integrated do not appropriately influence the agent's behavior, and (4) behavior is limited by the capacity of the model. In this paper we explore an alternative paradigm in which we train a network to map a dataset of past experiences to optimal behavior. Specifically, we augment an RL agent with a retrieval process (parameterized as a neural network) that has direct access to a dataset of experiences. This dataset can come from the agent's past experiences, expert demonstrations, or any other relevant source. The retrieval process is trained to retrieve information from the dataset that may be use

ful in the current context, to help the agent achieve its goal faster and more efficiently. We integrate our method into two different RL agents: an offline DQN agent and an online R2D2 agent. In offline multi-task problems, we show that the retrieval-augmented DQN agent avoids task interference and learns faster than the baseline DQN agent. On Atari, we show that retrieval-augmented R2D2 learns significantly faster than the baseline R2D2 agent and achieves higher scores. We run extensive ablations to measure the contributions of the components of our proposed method.

****

Scaling Densities For Improved Density Ratio Estimation

Akash Srivastava,Seungwook Han,Benjamin Rhodes,Kai Xu,Michael U. Gutmann

Estimating the discrepancy between two densities (p and q) is central to machine learning. Most frequently used methods for the quantification of this discrepancy capture it as a function of the ratio of the densities p/q. In practice, closed-form expressions for these densities or their ratio are rarely available. As such, estimating density ratios accurately using only samples from p and q is of high significance and has led to a flurry of recent work in this direction. Among these, binary classification based density ratio estimators have shown great promise and have been extremely successful in specialized domains. However, estimating the density ratio using a binary classifier, when the samples from the densities are well separated, remains challenging. In this work, we first show that the state-of-the-art solutions for such well-separated cases have limited applicability, may suffer from theoretical inconsistencies or lack formal guarantees and therefore perform poorly in the general case. We then present an alternative framework for density ratio estimation that is motivated by the scaled-Bregman divergence. Our proposal is to scale the densities p and q by another density m and estimate $\log p/q$ as $\log p/m - \log q/m$. We show that if the scaling measures are constructed such that they overlap with p and q, then a single multi-class logistic regression can be trained to accurately recover p/m and q/m on samples from p, q and m. We formally justify our method with the scaled-Bregman theorem and show that it does not suffer from the issues that plague the existing solutions.

****

Orchestrated Value Mapping for Reinforcement Learning

Mehdi Fatemi,Arash Tavakoli

We present a general convergent class of reinforcement learning algorithms that is founded on two distinct principles: (1) mapping value estimates to a different space using arbitrary functions from a broad class, and (2) linearly decomposing the reward signal into multiple channels. The first principle enables incorporating specific properties into the value estimator that can enhance learning. The second principle, on the other hand, allows for the value function to be represented as a composition of multiple utility functions. This can be leveraged for various purposes, e.g. dealing with highly varying reward scales, incorporating a priori knowledge about the sources of reward, and ensemble learning. Combining the two principles yields a general blueprint for instantiating convergent algorithms by orchestrating diverse mapping functions over multiple reward channels. This blueprint generalizes and subsumes algorithms such as Q-Learning, Log Q-Learning, and Q-Decomposition. In addition, our convergence proof for this general class relaxes certain required assumptions in some of these algorithms. Based on our theory, we discuss several interesting configurations as special cases. Finally, to illustrate the potential of the design space that our theory opens up, we instantiate a particular algorithm and evaluate its performance on the Atari suite.

****

Learning to Generalize across Domains on Single Test Samples

Zehao Xiao,Xiantong Zhen,Ling Shao,Cees G. M. Snoek

We strive to learn a model from a set of source domains that generalizes well to unseen target domains. The main challenge in such a domain generalization scenario is the unavailability of any target domain data during training, resulting i n the learned model not being explicitly adapted to the unseen target domains. We propose learning to generalize across domains on single test samples. We leverage a meta-learning paradigm to learn our model to acquire the ability of adaptation with single samples at training time so as to further adapt itself to each single test sample at test time. We formulate the adaptation to the single test sample as a variational Bayesian inference problem, which incorporates the test sample as a conditional into the generation of model parameters. The adaptation to each test sample requires only one feed-forward computation at test time without any fine-tuning or self-supervised training on additional data from the unseen domains. Extensive ablation studies demonstrate that our model learns the ability to adapt models to each single sample by mimicking domain shifts during training. Further, our model achieves at least comparable -- and often better -- performance than state-of-the-art methods on multiple benchmarks for domain generalization.

****

SOSP: Efficiently Capturing Global Correlations by Second-Order Structured Pruning

Manuel Nonnenmacher,Thomas Pfeil,Ingo Steinwart,David Reeb

Pruning neural networks reduces inference time and memory costs. On standard hardware, these benefits will be especially prominent if coarse-grained structures, like feature maps, are pruned. We devise two novel saliency-based methods for second-order structured pruning (SOSP) which include correlations among all structures and layers. Our main method SOSP-H employs an innovative second-order approximation, which enables saliency evaluations by fast Hessian-vector products. SOSP-H thereby scales like a first-order method despite taking into account the full Hessian. We validate SOSP-H by comparing it to our second method SOSP-I that uses a well-established Hessian approximation, and to numerous state-of-the-art methods. While SOSP-H performs on par or better in terms of accuracy, it has clear advantages in terms of scalability and efficiency. This allowed us to scale SOSP-H to large-scale vision tasks, even though it captures correlations across all layers of the network. To underscore the global nature of our pruning methods, we evaluate their performance not only by removing structures from a pretrained network, but also by detecting architectural bottlenecks. We show that our algorithms allow to systematically reveal architectural bottlenecks, which we then remove to further increase the accuracy of the networks.

****

The Importance of the Current Input in Sequence Modeling

Christian Oliva,Luis F. Lago-Fernandez

The last advances in sequence modeling are mainly based on deep learning approaches. The current state of the art involves the use of variations of the standard LSTM architecture, combined with several tricks that improve the final prediction rates of the trained neural networks. However, in some cases, these adaptations might be too much tuned to the particular problems being addressed. In this article, we show that a very simple idea, to add a direct connection between the input and the output, skipping the recurrent module, leads to an increase of the prediction accuracy in sequence modeling problems related to natural language processing. Experiments carried out on different problems show that the addition of this kind of connection to a recurrent network always improves the results, regardless of the architecture and training-specific details. When this idea is introduced into the models that lead the field, the resulting networks achieve a new state-of-the-art perplexity in language modeling problems.

****

REFACTOR: Learning to Extract Theorems from Proofs

Jin Peng Zhou,Yuhuai Wu,Qiyang Li,Roger Baker Grosse

Human mathematicians are often good at recognizing modular and reusable theorems that make complex mathematical results within reach. In this paper, we propose a novel method called theoREm-from-prooF extrACTOR (REFACTOR) for training neural networks to mimic this ability in formal mathematical theorem proving. We show on a set of unseen proofs, REFACTOR is able to extract 19.6% of the theorems that humans would use to write the proofs. When applying the model to the exist ing Metamath library, REFACTOR extracted 16 new theorems. With newly extracted theorems, we show that the existing proofs in the MetaMath database can be refactored. The new theorems are used very frequently after refactoring, with an average usage of 733.5 times, and help to shorten the proof lengths. Lastly, we demonstrate that the prover trained on the new-theorem refactored dataset proves relatively 14-30% more test theorems by frequently leveraging a diverse set of newly extracted theorems.

****

Prototype memory and attention mechanisms for few shot image generation

Tianqin Li,Zijie Li,Andrew Luo,Harold Rockwell,Amir Barati Farimani,Tai Sing Lee

Recent discoveries indicate that the neural codes in the primary visual cortex (V1) of macaque monkeys are complex, diverse and sparse. This leads us to ponder the computational advantages and functional role of these “grandmother cells." Here, we propose that such cells can serve as prototype memory priors that bias and shape the distributed feature processing within the image generation process in the brain. These memory prototypes are learned by momentum online clustering and are utilized via a memory-based attention operation, which we define as Memory Concept Attention (MoCA). To test our proposal, we show in a few-shot image generation task, that having a prototype memory during attention can improve image synthesis quality, learn interpretable visual concept clusters, as well as improve the robustness of the model. Interestingly, we also find that our attentional memory mechanism can implicitly modify the horizontal connections by updating the transformation into the prototype embedding space for self-attention. Insofar as GANs can be seen as plausible models for reasoning about the top-down synthesis in the analysis-by-synthesis loop of the hierarchical visual cortex, our findings demonstrate a plausible computational role for these “prototype concept" neurons in visual processing in the brain.

****

LatTe Flows: Latent Temporal Flows for Multivariate Sequence Analysis

Magda Amiridi,Gregory Darnell,Sean Jewell

We introduce Latent Temporal Flows (\emph{LatTe-Flows}), a method for probabilistic multivariate time-series analysis tailored for high dimensional systems whose temporal dynamics are driven by variations in a lower-dimensional discriminative subspace. We perform indirect learning from hidden traits of observed sequences by assuming that the random vector representing the data is generated from an unobserved low-dimensional latent vector. \emph{LatTe-Flows} jointly learns auto-encoder mappings to a latent space and learns the temporal distribution of lower-dimensional embeddings of input sequences. Since encoder networks retain only the essential information to generate a latent manifold, the temporal distribution transitions can be more efficiently uncovered by time conditioned Normalizing Flows. The learned latent effects can then be directly transferred into the observed space through the decoder network. We demonstrate that the proposed method significantly outperforms the state-of-the-art on multi-step forecasting benchmarks, while enjoying reduced computational complexity on several real-world datasets. We apply {\emph{LatTe-Flows}} to a challenging sensor-signal forecasting task, using multivariate time-series measurements collected by wearable devices, an increasingly relevant health application.

****

Relational Multi-Task Learning: Modeling Relations between Data and Tasks

Kaidi Cao,Jiaxuan You,Jure Leskovec

A key assumption in multi-task learning is that at the inference time the multi-task model only has access to a given data point but not to the data point’s labels from other tasks. This presents an opportunity to extend multi-task learning to utilize data point’s labels from other auxiliary tasks, and this way improves performance on the new task. Here we introduce a novel relational multi-task learning setting where we leverage data point labels from auxiliary tasks to make more accurate predictions on the new task. We develop MetaLink, where our key innovation is to build a knowledge graph that connects data points and tasks and thus allows us to leverage labels from auxiliary tasks. The knowledge graph cons ists of two types of nodes: (1) data nodes, where node features are data embeddings computed by the neural network, and (2) task nodes, with the last layer’s weights for each task as node features. The edges in this knowledge graph capture data-task relationships, and the edge label captures the label of a data point on a particular task. Under MetaLink, we reformulate the new task as a link label prediction problem between a data node and a task node. The MetaLink framework provides flexibility to model knowledge transfer from auxiliary task labels to the task of interest. We evaluate MetaLink on 6 benchmark datasets in both biochemical and vision domains. Experiments demonstrate that MetaLink can successfully utilize the relations among different tasks, outperforming the state-of-the-art methods under the proposed relational multi-task learning setting, with up to 27% improvement in ROC AUC.

****

CoBERL: Contrastive BERT for Reinforcement Learning

Andrea Banino,Adria Puigdomenech Badia,Jacob C Walker,Tim Scholtes,Jovana Mitrovic,Charles Blundell

Many reinforcement learning (RL) agents require a large amount of experience to solve tasks. We propose Contrastive BERT for RL (COBERL), an agent that combines a new contrastive loss and a hybrid LSTM-transformer architecture to tackle the challenge of improving data efficiency. COBERL enables efficient and robust learning from pixels across a wide variety of domains. We use bidirectional masked prediction in combination with a generalization of a recent contrastive method to learn better representations for RL, without the need of hand engineered data augmentations. We find that COBERL consistently improves data efficiency across the full Atari suite, a set of control tasks and a challenging 3D environment, and often it also increases final score performance.

****

Learning to Pool in Graph Neural Networks for Extrapolation

Jihoon Ko,Taehyung Kwon,Kijung Shin,Juho Lee

Graph neural networks (GNNs) are one of the most popular approaches to using deep learning on graph-structured data, and they have shown state-of-the-art performances on a variety of tasks. However, according to a recent study, a careful choice of pooling functions, which are used for the aggregation and readout operations in GNNs, is crucial for enabling GNNs to extrapolate. Without proper choices of pooling functions, which varies across tasks, GNNs completely fail to generalize to out-of-distribution data, while the number of possible choices grows exponentially with the number of layers. In this paper, we present GNP, a L^p norm-like pooling function that is trainable end-to-end for any given task. Notably, GNP generalizes most of the widely-used pooling functions. We verify experimentally that simply using GNP for every aggregation and readout operation enables GNNs to extrapolate well on many node-level, graph-level, and set-related tasks; and GNP sometimes performs even better than the best-performing choices among existing pooling functions.

****

Bandwidth-based Step-Sizes for Non-Convex Stochastic Optimization

Xiaoyu Wang,Mikael Johansson

Many popular learning-rate schedules for deep neural networks combine a decaying trend with local perturbations that attempt to escape saddle points and bad local minima. We derive convergence guarantees for bandwidth-based step-sizes, a general class of learning-rates that are allowed to vary in a banded region. This framework includes many popular cyclic and non-monotonic step-sizes for which no theoretical guarantees were previously known. We provide worst-case guarantees for SGD on smooth non-convex problems under several bandwidth-based step sizes, including stagewise $1/\sqrt{t}$ and the popular \emph{step-decay} (``constant and then drop by a constant’’), which is also shown to be optimal. Moreover, we show that its momentum variant converges as fast as SGD with the bandwidth-based step-decay step-size. Finally, we propose novel step-size schemes in the bandwidth-based family and verify their efficiency on several deep neural network trai

ning tasks.
****

Born Again Neural Rankers
Zhen Qin,Le Yan,Yi Tay,Honglei Zhuang,Xuanhui Wang,Michael Bendersky,Marc Najork

We introduce Born Again neural Rankers (BAR) in the Learning to Rank (LTR) setting, where student rankers, trained in the Knowledge Distillation (KD) framework, are parameterized identically to their teachers. Unlike the existing ranking distillation work, which pursues a good trade-off between performance and efficiency, BAR adapts the idea of Born Again Networks (BAN) to ranking problems and significantly improves ranking performance of students over the teacher rankers without increasing model capacity. By examining the key differences between ranking distillation and common distillation for classification problems, we find that the key success factors of BAR lie in (1) an appropriate teacher score transformation function, and (2) a novel listwise distillation framework, both are specifically designed for ranking problems and are rarely studied in the knowledge distillation literature. Using the state-of-the-art neural ranking structures, BAR is able to push the limits of neural rankers above a recent rigorous benchmark study, and significantly outperforms strong gradient boosted decision tree based models on 7 out of 9 key metrics, the first time in the literature. In addition to the strong empirical results, we give theoretical explanations on why listwise distillation is effective for neural rankers.
****

TPU-GAN: Learning temporal coherence from dynamic point cloud sequences
Zijie Li,Tianqin Li,Amir Barati Farimani

Point cloud sequence is an important data representation that provides flexible shape and motion information. Prior work demonstrates that incorporating scene flow information into loss can make model learn temporally coherent feature spaces. However, it is prohibitively expensive to acquire point correspondence information across frames in real-world environments. In this work, we propose a super-resolution generative adversarial network (GAN) for upsampling dynamic point cloud sequences, which does not require point correspondence annotation. Our model, Temporal Point cloud Upsampling GAN (TPU-GAN), can implicitly learn the underlying temporal coherence from point cloud sequence, which in turn guides the generator to produce temporally coherent output. In addition, we propose a learnable masking module to adapt upsampling ratio according to the point distribution. We conduct extensive experiments on point cloud sequences from two different domains: particles in the fluid dynamical system and human action scanned data. The quantitative and qualitative evaluation demonstrates the effectiveness of our method on upsampling tasks as well as learning temporal coherence from irregular point cloud sequences.
****

A First-Occupancy Representation for Reinforcement Learning
Ted Moskovitz,Spencer R Wilson,Maneesh Sahani

Both animals and artificial agents benefit from state representations that support rapid transfer of learning across tasks and which enable them to efficiently traverse their environments to reach rewarding states. The successor representation (SR), which measures the expected cumulative, discounted state occupancy under a fixed policy, enables efficient transfer to different reward structures in an otherwise constant Markovian environment and has been hypothesized to underlie aspects of biological behavior and neural activity. However, in the real world, rewards may only be available for consumption once, may shift location, or agents may simply aim to reach goal states as rapidly as possible without the constraint of artificially imposed task horizons. In such cases, the most behaviorally-relevant representation would carry information about when the agent was likely to first reach states of interest, rather than how often it should expect to visit them over a potentially infinite time span. To reflect such demands, we introduce the first-occupancy representation (FR), which measures the expected temporal discount to the first time a state is accessed. We demonstrate that the FR facilitates exploration, the selection of efficient paths to desired states, allows the agent, under certain conditions, to plan provably optimal trajectori es defined by a sequence of subgoals, and induces similar behavior to animals avoiding threatening stimuli.

****

Deep ReLU Networks Preserve Expected Length

Boris Hanin,Ryan S Jeong,David Rolnick

Assessing the complexity of functions computed by a neural network helps us understand how the network will learn and generalize. One natural measure of complexity is how the network distorts length - if the network takes a unit-length curve as input, what is the length of the resulting curve of outputs? It has been widely believed that this length grows exponentially in network depth. We prove that in fact this is not the case: the expected length distortion does not grow with depth, and indeed shrinks slightly, for ReLU networks with standard random initialization. We also generalize this result by proving upper bounds both for higher moments of the length distortion and for the distortion of higher-dimensional volumes. These theoretical results are corroborated by our experiments.

****

Phenomenology of Double Descent in Finite-Width Neural Networks

Sidak Pal Singh,Aurelien Lucchi,Thomas Hofmann,Bernhard Schölkopf

`Double descent' delineates the generalization behaviour of models depending on the regime they belong to: under- or over-parameterized. The current theoretical understanding behind the occurrence of this phenomenon is primarily based on linear and kernel regression models --- with informal parallels to neural networks via the Neural Tangent Kernel. Therefore such analyses do not adequately capture the mechanisms behind double descent in finite-width neural networks, as well as, disregard crucial components --- such as the choice of the loss function. We address these shortcomings by leveraging influence functions in order to derive suitable expressions of the population loss and its lower bound, while imposing minimal assumptions on the form of the parametric model. Our derived bounds bear an intimate connection with the spectrum of the Hessian at the optimum, and importantly, exhibit a double descent behaviour at the interpolation threshold. Building on our analysis, we further investigate how the loss function affects double descent --- and thus uncover interesting properties of neural networks and their Hessian spectra near the interpolation threshold.

****

Optimal Transport for Causal Discovery

Ruibo Tu,Kun Zhang,Hedvig Kjellstrom,Cheng Zhang

To determine causal relationships between two variables, approaches based on Functional Causal Models (FCMs) have been proposed by properly restricting model classes; however, the performance is sensitive to the model assumptions, which makes it difficult to use. In this paper, we provide a novel dynamical-system view of FCMs and propose a new framework for identifying causal direction in the bivariate case. We first show the connection between FCMs and optimal transport, and then study optimal transport under the constraints of FCMs. Furthermore, by exploiting the dynamical interpretation of optimal transport under the FCM constraints, we determine the corresponding underlying dynamical process of the static cause-effect pair data. It provides a new dimension for describing static causal discovery tasks while enjoying more freedom for modeling the quantitative causal influences. In particular, we show that Additive Noise Models (ANMs) correspond to volume-preserving pressureless flows. Consequently, based on their velocity field divergence, we introduce a criterion for determining causal direction. With this criterion, we propose a novel optimal transport-based algorithm for ANMs which is robust to the choice of models and extend it to post-nonlinear models. Our method demonstrated state-of-the-art results on both synthetic and causal discovery benchmark datasets.

****

How Attentive are Graph Attention Networks?

Shaked Brody,Uri Alon,Eran Yahav

Graph Attention Networks (GATs) are one of the most popular GNN architectures and are considered as the state-of-the-art architecture for representation learning with graphs. In GAT, every node attends to its neighbors given its own represe ntation as the query.

However, in this paper we show that GAT computes a very limited kind of attention: the ranking of the attention scores is unconditioned on the query node. We formally define this restricted kind of attention as static attention and distinguish it from a strictly more expressive dynamic attention.

Because GATs use a static attention mechanism, there are simple graph problems that GAT cannot express: in a controlled problem, we show that static attention hinders GAT from even fitting the training data.

To remove this limitation, we introduce a simple fix by modifying the order of operations and propose GATv2: a dynamic graph attention variant that is strictly more expressive than GAT. We perform an extensive evaluation and show that GATv2 outperforms GAT across 12 OGB and other benchmarks while we match their parametric costs.

Our code is available at https://github.com/tech-srl/how_attentive_are_gats . GATv2 is available as part of the PyTorch Geometric library, the Deep Graph Library, and the TensorFlow GNN library.

****

Linear Convergence of SGD on Overparametrized Shallow Neural Networks

Paul Rolland,Ali Ramezani-Kebrya,ChaeHwan Song,Fabian Latorre,Volkan Cevher

Despite the non-convex landscape, first-order methods can be shown to reach global minima when training overparameterized neural networks, where the number of parameters far exceed the number of training data. In this work, we prove linear convergence of stochastic gradient descent when training a two-layer neural network with smooth activations. While the existing theory either requires a high degree of overparameterization or non-standard initialization and training strategies, e.g., training only a single layer, we show that a subquadratic scaling on the width is sufficient under standard initialization and training both layers simultaneously if the minibatch size is sufficiently large and it also grows with the number of training examples. Via the batch size, our results interpolate between the state-of-the-art subquadratic results for gradient descent and the quadratic results in the worst case.

****

Localized Persistent Homologies for more Effective Deep Learning

Doruk Oner,Adélie Garin,Mateusz Kozinski,Kathryn Hess,Pascal Fua

Persistent Homologies have been successfully used to increase the performance of deep networks trained to detect curvilinear structures and to improve the topological quality of the results. However, existing methods are very global and ignore the location of topological features. In this paper, we introduce an approach that relies on a new filtration function to account for location during network training. We demonstrate experimentally on 2D images of roads and 3D image stacks of neural processes that networks trained in this manner are better at recovering the topology of the curvilinear structures they extract.

****

Estimating and Penalizing Induced Preference Shifts in Recommender Systems

Micah Carroll,Dylan Hadfield-Menell,Stuart Russell,Anca Dragan

The actions that a recommender system (RS) takes -- the content it exposes users to -- influence the preferences users have over what content they want. Therefore, when an RS designer chooses which system to deploy, they are implicitly \emph{choosing how to shift} or influence user preferences. Even more, if the RS is trained via long-horizon optimization (e.g. reinforcement learning), it will have incentives to manipulate preferences, i.e to shift them so they are more easy to satisfy, and thus conducive to higher reward. While some work has argued for making systems myopic to avoid this issue, myopic systems can still influence preferences in undesirable ways. In this work, we argue that we need to enable system designers to \textit{estimate} the shifts an RS \emph{would} induce; \textit{evaluate}, before deployment, whether the shifts are undesirable; and even \textit{actively optimize} to avoid such shifts. These steps involve two challenging ingredients: \emph{estimation} requires the ability to anticipate how hypothetical policies would influence user preferences if deployed -- we do this by training a user predictive model that implicitly contains their preference dynamics f rom historical user interaction data; \emph{evaluation} and \emph{optimization} additionally require metrics to assess whether such influences are manipulative or otherwise unwanted -- we introduce the notion of “safe shifts”, that define a trust region within which behavior is believed to be safe. We show that recommender systems that optimize for staying in the trust region can avoid manipulative behaviors (e.g., changing preferences in ways that make users more predictable), while still generating engagement.

****

SGORNN: Combining Scalar Gates and Orthogonal Constraints in Recurrent Networks
William Keith Taylor-Melanson,Martha Dais Ferreira,Stan Matwin

Recurrent Neural Network (RNN) models have been applied in different domains, producing high accuracies on time-dependent data. However, RNNs have long suffered from exploding gradients during training, mainly due to their recurrent process. In this context, we propose a variant of the scalar gated FastRNN architecture, called Scalar Gated Orthogonal Recurrent Neural Networks (SGORNN). SGORNN utilizes orthogonal linear transformations at the recurrent step. In our experiments, SGORNN forms its recurrent weights through a strategy inspired by Volume Preserving RNNs (VPRNN), though our architecture allows the use of any orthogonal constraint mechanism. We present a simple constraint on the scalar gates of SGORNN, which is easily enforced at training time to provide a theoretical generalization ability for SGORNN similar to that of FastRNN. Our constraint is further motivated by success in experimental settings. Next, we provide bounds on the gradients of SGORNN, to show the impossibility of (exponentially) exploding gradients. Our experimental results on the addition problem confirm that our combination of orthogonal and scalar gated RNNs are able to outperform both predecessor models on long sequences using only a single RNN cell. We further evaluate SGORNN on the HAR-2 classification task, where it improves slightly upon the accuracy of both FastRNN and VPRNN using far fewer parameters than FastRNN. Finally, we evaluate SGORNN on the Penn Treebank word-level language modelling task, where it again outperforms its predecessor architectures. Overall, this architecture shows higher representation capacity than VPRNN, suffers from less overfitting than the other two models in our experiments, benefits from a decrease in parameter count, and alleviates exploding gradients when compared with FastRNN on the addition problem.

****

Global Convergence and Stability of Stochastic Gradient Descent
Vivak Patel,Bowen Tian,Shushu Zhang

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 make substantial progress on this shortcoming. 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.

****

Contrastive Learning of 3D Shape Descriptor with Dynamic Adversarial Views
Shuaihang Yuan,Yi Fang

View-based deep learning models have shown the capability to learn 3D shape descriptors with superior performance on 3D shape recognition, classification, and retrieval. Most popular techniques often leverage the class label to train deep neural networks under supervision to learn to extract 3D deep representation by aggregating information from a static and pre-selected set of different views used for all shapes. Those approaches, however, often face challenges posed by the requirement of a large amount of annotated training data and the lack of a mecha nism for the adaptive selection of shape-instance-dependent views towards the learning of more informative 3D shape representation. This paper addresses those two challenging issues by introducing the concept of adversarial views and developing a new mechanism to generate views for adversarial training of a self-supervised contrastive model for 3D shape descriptor, denoted as CoLAV. In particular, compared to the recent advances in multi-view approaches, our proposed CoLAV gains advantages by leveraging the contrastive learning techniques for self-supervised learning of 3D shape representations without the need for labeled data. In addition, CoLAV introduces a novel mechanism for the dynamic generation of shape-instance-dependent adversarial views as positive pairs to adversarially train robust contrastive learning models towards the learning of more informative 3D shape representation. Comprehensive experimental results on 3D shape classification demonstrate that the 3D shape descriptor learned by CoLAV exhibits superior performance for 3D shape recognition over other state-of-the-art techniques, even though CoLAV is completely self-trained with unlabeled 3D datasets (e.g., ModelNet40).

****

A Distributional Robustness Perspective on Adversarial Training with the ∞-Wasserstein Distance

Chiara Regniez,Gauthier Gidel,Hugo berard

While ML tools are becoming increasingly used in industrial applications, adversarial examples remain a critical flaw of neural networks. These imperceptible perturbations of natural inputs are, on average, misclassified by most of the state-of-the-art classifiers. By slightly modifying each data point, the attacker is creating a new distribution of inputs for the classifier. In this work, we consider the adversarial examples distribution as a tiny shift of the original distribution. We thus propose to address the problem of adversarial training (AT) within the framework of distributional robustness optimization (DRO). We show a formal connection between our formulation and optimal transport by relaxing AT into DRO problem with an ∞-Wasserstein constraint. This connection motivates using an entropic regularizer-- a standard tool in optimal transport--- for our problem. We then prove the existence and uniqueness of an optimal regularized distribution of adversarial examples against a class of classifier (e.g., a given architecture) that we eventually use to robustly train a classifier. Using these theoretical insights, we propose to use Langevin Monte Carlo to sample from this optimal distribution of adversarial examples and train robust classifiers outperforming the standard baseline and providing a speed-up of respectively $\times 200$ for MNIST and $\times8$ for CIFAR-10.

****

Learning Transferable Reward for Query Object Localization with Policy Adaptation

Tingfeng Li,Shaobo Han,Martin Renqiang Min,Dimitris N. Metaxas

We propose a reinforcement learning based approach to query object localization, for which an agent is trained to localize objects of interest specified by a small exemplary set. We learn a transferable reward signal formulated using the exemplary set by ordinal metric learning. Our proposed method enables test-time policy adaptation to new environments where the reward signals are not readily available, and outperforms fine-tuning approaches that are limited to annotated images. In addition, the transferable reward allows repurposing the trained agent from one specific class to another class. Experiments on corrupted MNIST, CU-Birds, and COCO datasets demonstrate the effectiveness of our approach.

****

Unifying Categorical Models by Explicit Disentanglement of the Labels' Generative Factors

Luca Pinchetti,Lei Sha,Thomas Lukasiewicz

In most machine learning tasks, the datasets are mainly annotated by categorical labels. For example, in emotion recognition, most datasets rely only on categorical labels, such as ``happy'' and ``sad''. Usually, different datasets use different labelling systems (e.g., different number of categories and different names), even when describing the same data attributes. As a consequence, only a smal l subset of all the available datasets can be used for any supervised learning task, since the labelling systems used in the training data are not compatible with each other.

In this paper, we propose a \emph{multi-type continuous disentanglement variational autoencoder} to address this problem by identifying and disentangling the true dimensional generative factors that determine each categorical label. By doing so, it is possible to merge multiple datasets based on different categorical models by projecting the data points into a unified latent space.

The experiments performed on synthetic datasets show a perfect correlation between the disentangled latent values and the true generative factors. Also, by observing the displacement of each label's explicit distributions, we noticed that the encoded space is a simple affine transformation of the generative factors' space. As the latent structure can be autonomously learnt by the model, and each label can be explicitly decomposed in its generative factors, this framework is very promising for further exploring explainability in new and existing neural networks architectures.

****

Communicating via Markov Decision Processes

Samuel Sokota,Christian Schroeder de Witt,Maximilian Igl,Luisa M Zintgraf,Philip Torr,J Zico Kolter,Shimon Whiteson,Jakob Nicolaus Foerster

We consider the problem of communicating exogenous information by means of Markov decision process trajectories. This setting, which we call a Markov coding game (MCG), generalizes both source coding and a large class of referential games. MCGs also isolate a problem that is important in decentralized control settings in which cheap-talk is not available---namely, they require balancing communication with the associated cost of communicating. We contribute a theoretically grounded approach to MCGs based on maximum entropy reinforcement learning and minimum entropy coupling that we call greedy minimum entropy coupling (GME). We show both that GME is able to outperform a relevant baseline on small MCGs and that GME is able to scale efficiently to extremely large MCGs. To the latter point, we demonstrate that GME is able to losslessly communicate binary images via trajectories of Cartpole and Pong, while simultaneously achieving the maximal or near maximal expected returns, and that it is even capable of performing well in the presence of actuator noise.

****

A Simple Approach to Adversarial Robustness in Few-shot Image Classification

Akshayvarun Subramanya,Hamed Pirsiavash

Few-shot image classification, where the goal is to generalize to tasks with limited labeled data, has seen great progress over the years. However, the classifiers are vulnerable to adversarial examples, posing a question regarding their generalization capabilities. Recent works have tried to combine meta-learning approaches with adversarial training to improve the robustness of few-shot classifiers. We show that a simple transfer-learning based approach can be used to train adversarially robust few-shot classifiers. We also present a method for novel classification task based on calibrating the centroid of the few-shot category towards the base classes. We show that standard adversarial training on base categories along with centroid-based classifier in the novel categories, outperforms or is on-par with state-of-the-art advanced methods on standard benchmarks such as Mini-ImageNet, CIFAR-FS and CUB datasets. Our method is simple and easy to scale, and with little effort can lead to robust few-shot classifiers.

****

Blur Is an Ensemble: Spatial Smoothings to Improve Accuracy, Uncertainty, and Robustness

Namuk Park,Songkuk Kim

Bayesian neural networks (BNNs) have shown success in the areas of uncertainty estimation and robustness. However, a crucial challenge prohibits their use in practice. Bayesian NNs require a large number of predictions to produce reliable results, leading to a significant increase in computational cost. To alleviate this issue, we propose spatial smoothing, a method that ensembles neighboring feature map points of CNNs. By simply adding a few blur layers to the models, we emp irically show that spatial smoothing improves accuracy, uncertainty estimation, and robustness of BNNs across a whole range of ensemble sizes. In particular, BNNs incorporating spatial smoothing achieve high predictive performance merely with a handful of ensembles. Moreover, this method also can be applied to canonical deterministic neural networks to improve the performances. A number of evidences suggest that the improvements can be attributed to the stabilized feature maps and the flattening of the loss landscape. In addition, we provide a fundamental explanation for prior works—namely, global average pooling, pre-activation, and ReLU6—by addressing them as special cases of spatial smoothing. These not only enhance accuracy, but also improve uncertainty estimation and robustness by making the loss landscape smoother in the same manner as spatial smoothing.

****

GroupBERT: Enhanced Transformer Architecture with Efficient Grouped Structures
Ivan Chelombiev,Daniel Justus,Douglas Orr,Anastasia S. D. Dietrich,Frithjof Gressmann,Alexandros Koliousis,Carlo Luschi
Attention based language models have become a critical component in state-of-the-art natural language processing systems. However, these models have significant computational requirements, due to long training times, dense operations and large parameter count. In this work we demonstrate a set of modifications to the structure of a Transformer layer, producing a more efficient architecture. First, we rely on grouped transformations to reduce the computational cost of dense feed-forward layers, while preserving the expressivity of the model . Secondly, we add a grouped convolution module to complement the self-attention module, decoupling the learning of local and global interactions. We apply the resulting architecture to language representation learning and demonstrate its superior performance compared to BERT models of different scales. We further highlight its improved efficiency, both in terms of floating-point operations (FLOPs) and time-to-train.

****

CKConv: Continuous Kernel Convolution For Sequential Data
David W. Romero,Anna Kuzina,Erik J Bekkers,Jakub Mikolaj Tomczak,Mark Hoogendoorn
Conventional neural architectures for sequential data present important limitations. Recurrent neural networks suffer from exploding and vanishing gradients, small effective memory horizons, and must be trained sequentially. Convolutional neural networks cannot handle sequences of unknown size and their memory horizon must be defined a priori. In this work, we show that these problems can be solved by formulating the convolutional kernels of CNNs as continuous functions. The resulting Continuous Kernel Convolution (CKConv) handles arbitrarily long sequences in a parallel manner, within a single operation, and without relying on any form of recurrence. We show that Continuous Kernel Convolutional Networks (CKCNNs) obtain state-of-the-art results in multiple datasets, e.g., permuted MNIST, and, thanks to their continuous nature, are able to handle non-uniformly sampled datasets and irregularly-sampled data natively. CKCNNs match or perform better than neural ODEs designed for these purposes in a faster and simpler manner.

****

3D Meta-Registration: Meta-learning 3D Point Cloud Registration Functions
Yu Hao,Yi Fang
Learning robust 3D point cloud registration functions with deep neural networks has emerged as a powerful paradigm in recent years, offering promising performance in producing spatial geometric transformations for each pair of 3D point clouds. However, 3D point cloud registration functions are often generalized from extensive training over a large volume of data to learn the ability to predict the desired geometric transformation to register 3D point clouds. Generalizing across 3D point cloud registration functions requires robust learning of priors over the respective function space and enables consistent registration in presence of significant 3D structure variations. In this paper, we proposed to formalize the learning of a 3D point cloud registration function space as a meta-learning problem, aiming to predict a 3D registration model that can be quickly adapted to

new point clouds with no or limited training data. Specifically, we define each task as the learning of the 3D registration function which takes points in 3D space as input and predicts the geometric transformation that aligns the source point cloud with the target one. Also, we introduce an auxiliary deep neural network named 3D registration meta-learner that is trained to predict the prior over the respective 3D registration function space. After training, the 3D registration meta-learner, which is trained with the distribution of 3D registration function space, is able to uniquely parameterize the 3D registration function with optimal initialization to rapidly adapt to new registration tasks. We tested our model on the synthesized dataset ModelNet and FlyingThings3D, as well as real-world dataset KITTI. Experimental results demonstrate that 3D Meta-Registration achieves superior performance over other previous techniques (e.g. FlowNet3D).

****

On Bridging Generic and Personalized Federated Learning for Image Classification

Hong-You Chen,Wei-Lun Chao

Federated learning is promising for its capability to collaboratively train models with multiple clients without accessing their data, but vulnerable when clients' data distributions diverge from each other. This divergence further leads to a dilemma: "Should we prioritize the learned model's generic performance (for future use at the server) or its personalized performance (for each client)?" These two, seemingly competing goals have divided the community to focus on one or the other, yet in this paper we show that it is possible to approach both at the same time. Concretely, we propose a novel federated learning framework that explicitly decouples a model's dual duties with two prediction tasks. On the one hand, we introduce a family of losses that are robust to non-identical class distributions, enabling clients to train a generic predictor with a consistent objective across them. On the other hand, we formulate the personalized predictor as a lightweight adaptive module that is learned to minimize each client's empirical risk on top of the generic predictor. With this two-loss, two-predictor framework which we name Federated Robust Decoupling (Fed-RoD), the learned model can simultaneously achieve state-of-the-art generic and personalized performance, essentially bridging the two tasks.

****

Towards Empirical Sandwich Bounds on the Rate-Distortion Function

Yibo Yang,Stephan Mandt

Rate-distortion (R-D) function, a key quantity in information theory, characterizes the fundamental limit of how much a data source can be compressed subject to a fidelity criterion, by any compression algorithm. As researchers push for ever-improving compression performance, establishing the R-D function of a given data source is not only of scientific interest, but also reveals the possible room for improvement in existing compression algorithms. Previous work on this problem relied on distributional assumptions on the data source (Gibson, 2017) or only applied to discrete data (Blahut, 1972; Arimoto, 1972). By contrast, this paper makes the first attempt at an algorithm for sandwiching the R-D function of a general (not necessarily discrete) source requiring only i.i.d. data samples. We estimate R-D sandwich bounds for a variety of artificial and real-world data sources, in settings far beyond the feasibility of any known method, and shed light on the optimality of neural data compression (Ballé et al., 2021; Yang et al., 2022). Our R-D upper bound on natural images indicates theoretical room for improving state-of-the-art image compression methods by at least one dB in PSNR at various bitrates. Our data and code can be found at https://github.com/mandt-lab/RD-sandwich.

****

Learning to Generalize Compositionally by Transferring Across Semantic Parsing Tasks

Wang Zhu,Peter Shaw,Tal Linzen,Fei Sha

Neural network models often generalize poorly to mismatched domains or distributions. In NLP, this issue arises in particular when models are expected to generalize compositionally, that is, to novel combinations of familiar words and constructions. We investigate learning representations that facilitate transfer learn ing from one compositional task to another: the representation and the task-specific layers of the models are strategically trained differently on a pre-finetuning task such that they generalize well on mismatched splits that require compositionality. We apply this method to semantic parsing, using three very different datasets, COGS, GeoQuery and SCAN, used alternately as the pre-finetuning and target task. Our method significantly improves compositional generalization over baselines on the test set of the target task, which is held out during fine-tuning. Ablation studies characterize the utility of the major steps in the proposed algorithm and support our hypothesis.

****

Pareto Policy Adaptation

Panagiotis Kyriakis,Jyotirmoy Deshmukh,Paul Bogdan

We present a policy gradient method for Multi-Objective Reinforcement Learning under unknown, linear preferences. By enforcing Pareto stationarity, a first-order condition for Pareto optimality, we are able to design a simple policy gradient algorithm that approximates the Pareto front and infers the unknown preferences. Our method relies on a projected gradient descent solver that identifies common ascent directions for all objectives. Leveraging the solution of that solver, we introduce Pareto Policy Adaptation (PPA), a loss function that adapts the policy to be optimal with respect to any distribution over preferences. PPA uses implicit differentiation to back-propagate the loss gradient bypassing the operations of the projected gradient descent solver. Our approach is straightforward, easy to implement and can be used with all existing policy gradient and actor-critic methods. We evaluate our method in a series of reinforcement learning tasks

****

A Sampling-Free Approximation of Gaussian Variational Auto-Encoders

Felix Petersen,Christian Borgelt,Hilde Kuehne,Oliver Deussen

We propose a sampling-free approximate formulation of Gaussian variational auto-encoders. Instead of computing the loss via stochastic sampling, we propagate the Gaussian distributions from the latent space into the output space. As computing the exact likelihood probability is intractable, we propose to locally approximate the decoder network by its Taylor series. We demonstrate that this approximation allows us to approximate the Gaussian variational auto-encoder training objective in closed form. We evaluate the proposed method on the CelebA, the 3D Chairs, and the MNIST data sets. We find that our sampling-free approximation performs better than its sampling counterpart on the Frechet inception distance and on par on the estimated marginal likelihood.

****

FEVERLESS: Fast and Secure Vertical Federated Learning based on XGBoost for Decentralized Labels

Rui Wang,O■uzhan Ersoy,Hangyu Zhu,Yaochu Jin,Kaitai Liang

Vertical Federated Learning (VFL) enables multiple clients to collaboratively train a global model over vertically partitioned data without revealing private local information. Tree-based models, like XGBoost and LightGBM, have been widely used in VFL to enhance the interpretation and efficiency of training. However, there is a fundamental lack of research on how to conduct VFL securely over distributed labels. This work is the first to fill this gap by designing a novel protocol, called FEVERLESS, based on XGBoost. FEVERLESS leverages secure aggregation via information masking technique and global differential privacy provided by a fairly and randomly selected noise leader to prevent private information from being leaked in the training process. Furthermore, it provides label and data privacy against honest-but-curious adversary even in the case of collusion of $n-2$ out of n clients. We present a comprehensive security and efficiency analysis for our design, and the empirical experiment results demonstrate that FEVERLESS is fast and secure. In particular, it outperforms the solution based on additive homomorphic encryption in runtime cost and provides better accuracy than the local differential privacy approach.

****

Fair Normalizing Flows

Mislav Balunovic,Anian Ruoss,Martin Vechev

Fair representation learning is an attractive approach that promises fairness of downstream predictors by encoding sensitive data. Unfortunately, recent work has shown that strong adversarial predictors can still exhibit unfairness by recovering sensitive attributes from these representations. In this work, we present Fair Normalizing Flows (FNF), a new approach offering more rigorous fairness guarantees for learned representations. Specifically, we consider a practical setting where we can estimate the probability density for sensitive groups. The key idea is to model the encoder as a normalizing flow trained to minimize the statistical distance between the latent representations of different groups. The main advantage of FNF is that its exact likelihood computation allows us to obtain guarantees on the maximum unfairness of any potentially adversarial downstream predictor. We experimentally demonstrate the effectiveness of FNF in enforcing various group fairness notions, as well as other attractive properties such as interpretability and transfer learning, on a variety of challenging real-world datasets.

****

The Convex Geometry of Backpropagation: Neural Network Gradient Flows Converge to Extreme Points of the Dual Convex Program

Yifei Wang,Mert Pilanci

We study non-convex subgradient flows for training two-layer ReLU neural networks from a convex geometry and duality perspective. We characterize the implicit bias of unregularized non-convex gradient flow as convex regularization of an equivalent convex model. We then show that the limit points of non-convex subgradient flows can be identified via primal-dual correspondence in this convex optimization problem. Moreover, we derive a sufficient condition on the dual variables which ensures that the stationary points of the non-convex objective are the KKT points of the convex objective, thus proving convergence of non-convex gradient flows to the global optimum. For a class of regular training data distributions such as orthogonal separable data, we show that this sufficient condition holds. Therefore, non-convex gradient flows in fact converge to optimal solutions of a convex optimization problem. We present numerical results verifying the predictions of our theory for non-convex subgradient descent.

****

Adaptive Wavelet Transformer Network for 3D Shape Representation Learning

Hao Huang,Yi Fang

We present a novel method for 3D shape representation learning using multi-scale wavelet decomposition. Previous works often decompose 3D shapes into complementary components in spatial domain at a single scale. In this work, we study to decompose 3D shapes into sub-bands components in frequency domain at multiple scales, resulting in a hierarchical decomposition tree in a principled manner rooted in multi-resolution wavelet analysis. Specifically, we propose Adaptive Wavelet Transformer Network (AWT-Net) that firstly generates approximation or detail wavelet coefficients per point, classifying each point into high or low sub-bands components, using lifting scheme at multiple scales recursively and hierarchically. Then, AWT-Net exploits Transformer to enhance the original shape features by querying and fusing features from different but integrated sub-bands. The wavelet coefficients can be learned without direct supervision on coefficients, and AWT-Net is fully differentiable and can be learned in an end-to-end fashion. Extensive experiments demonstrate that AWT-Net achieves competitive performance on 3D shape classification and segmentation benchmarks.

****

Semantically Controllable Generation of Physical Scenes with Explicit Knowledge

Wenhao Ding,Bo Li,Ji Eun Kim,Ding Zhao

Deep Generative Models (DGMs) are known for their superior capability in generating realistic data. Extending purely data-driven approaches, recent specialized DGMs satisfy additional controllable requirements such as embedding a traffic sign in a driving scene by manipulating patterns implicitly in the neuron or feature level. In this paper, we introduce a novel method to incorporate domain knowledge explicitly in the generation process to achieve the semantically controllable generation of physical scenes. We first categorize our knowledge into two typ es, the property of objects and the relationship among objects, to be consistent with the composition of natural scenes. We then propose a tree-structured generative model to learn hierarchical scene representation, whose nodes and edges naturally corresponded to the two types of knowledge, respectively. Consequently, explicit knowledge integration enables semantically controllable generation by imposing semantic rules on the properties of nodes and edges in the tree structure. We construct a synthetic example to illustrate the controllability and explainability of our method in a succinct setting. We further extend the synthetic example to realistic environments for autonomous vehicles and conduct extensive experiments: our method efficiently identifies adversarial physical scenes against different state-of-the-art 3D point cloud segmentation models and satisfies the traffic rules specified as the explicit knowledge.

****

LEARNING DISTRIBUTIONS GENERATED BY SINGLE-LAYER RELU NETWORKS IN THE PRESENCE OF ARBITRARY OUTLIERS

Saikiran Bulusu,Geethu Joseph,M. Cenk Gursoy,Pramod Varshney

We consider a set of data samples such that a constant fraction of the samples are arbitrary outliers and the rest are the output samples of a single-layer neural network (NN) with rectified linear unit (ReLU) activation. The goal of this paper is to estimate the parameters (weight matrix and bias vector) of the NN assuming the bias vector to be non-negative. Our proposed method is a two-step algorithm. We first estimate the norms of the rows of the weight matrix and the bias vector using the gradient descent algorithm. Here, we also incorporate either the median or the trimmed mean based filters to mitigate the effect of the arbitrary outliers. Next, we estimate the angles between any two row vectors of the weight matrix. Combining the estimates of the norms and the angles, we obtain the final estimate of the weight matrix. Further, we prove that ${O}(\frac{1}{\epsilon p^4}\log\frac{d}{\delta})$ samples are sufficient for our algorithm to estimate the NN parameters within an error of ϵ with probability $1-\delta$ when the probability of a sample being uncorrupted is p and the problem dimension is d. Our theoretical and simulation results provide insights on how the estimation of the NN parameters depends on the probability of a sample being uncorrupted, the number of samples, and the problem dimension.

****

A Joint Subspace View to Convolutional Neural Networks

Ze Wang,Xiuyuan Cheng,Guillermo Sapiro,Qiang Qiu

Motivated by the intuition that important image regions remain important across different layers and scales in a CNN, we propose in this paper a joint subspace view to convolutional filters across network layers. When we construct for each layer a filter subspace by decomposing convolutional filters over a small set of layer-specific filter atoms, we observe a low-rank structure within subspace coefficients across layers. The above observation matches widely-known cross-layer filter correlation and redundancy. Thus, we propose to jointly model filter subspace across different layers by enforcing cross-layer shared subspace coefficients. In other words, a CNN is now reduced to layers of filter atoms, typically a few hundred of parameters per layer, with a common block of subspace coefficients shared across layers. We further show that such subspace coefficient sharing can be easily extended to other network sub-structures, from sharing across the entire network to sharing within filter groups in a layer. While significantly reducing the parameter redundancy of a wide range of network architectures, the proposed joint subspace view also preserves the expressiveness of CNNs, and brings many additional advantages, such as easy model adaptation and better interpretation. We support our findings with extensive empirical evidence.

****

Depth Without the Magic: Inductive Bias of Natural Gradient Descent

Anna Kerekes,Anna Mészáros,Ferenc Huszár

In gradient descent, changing how we parametrize the model can lead to drastically different optimization trajectories, giving rise a surprising range of meaningful inductive biases: identifying sparse classifiers or reconstructing low-rank matrices without explicit regularization. This implicit regularization has been hypothesised to be a contributing factor to good generalization in deep learning. However, natural gradient descent is approximately invariant to reparameterization, it always follows the same trajectory and finds the same optimum. The question naturally arises: What happens if we eliminate the role of parameterization, which solution will be found, what new properties occur? We characterize the behaviour of natural gradient flow in deep linear networks for separable classification under logistic loss and deep matrix factorization. Some of our findings extend to nonlinear neural networks with sufficient but finite over-parametrization. We demonstrate that there exist learning problems where natural gradient descent fails to generalize, while gradient descent with the right architecture peforms well.

****

k-Mixup Regularization for Deep Learning via Optimal Transport

Kristjan Greenewald,Anming Gu,Mikhail Yurochkin,Justin Solomon,Edward Chien

Mixup is a popular regularization technique for training deep neural networks that can improve generalization and increase adversarial robustness. It perturbs input training data in the direction of other randomly-chosen instances in the training set. To better leverage the structure of the data, we extend mixup to k-mixup by perturbing k-batches of training points in the direction of other k-batches using displacement interpolation, i.e. interpolation under the Wasserstein metric. We demonstrate theoretically and in simulations that k-mixup preserves cluster and manifold structures, and we extend theory studying the efficacy of standard mixup to the k-mixup case. Our empirical results show that training with k-mixup further improves generalization and robustness across several network architectures and benchmark datasets of differing modalities.

****

On the Convergence of mSGD and AdaGrad for Stochastic Optimization

ruinan Jin,Yu Xing,Xingkang He

As one of the most fundamental stochastic optimization algorithms, stochastic gradient descent (SGD) has been intensively developed and extensively applied in machine learning in the past decade. There have been some modified SGD-type algorithms, which outperform the SGD in many competitions and applications in terms of convergence rate and accuracy, such as momentum-based SGD (mSGD) and adaptive gradient algorithm (AdaGrad). Despite these empirical successes, the theoretical properties of these algorithms have not been well established due to technical difficulties. With this motivation, we focus on convergence analysis of mSGD and AdaGrad for any smooth (possibly non-convex) loss functions in stochastic optimization. First, we prove that the iterates of mSGD are asymptotically convergent to a connected set of stationary points with probability one, which is more general than existing works on subsequence convergence or convergence of time averages. Moreover, we prove that the loss function of mSGD decays at a certain rate faster than that of SGD. In addition, we prove the iterates of AdaGrad are asymptotically convergent to a connected set of stationary points with probability one. Also, this result extends the results from the literature on subsequence convergence and the convergence of time averages. Despite the generality of the above convergence results, we have relaxed some assumptions of gradient noises, convexity of loss functions, as well as boundedness of iterates.

****

DCoM: A Deep Column Mapper for Semantic Data Type Detection

Subhadip Maji,Swapna sourav Rout,Sudeep Choudhary

Detection of semantic data types is a very crucial task in data science for automated data cleaning, schema matching, data discovery, semantic data type normalization and sensitive data identification. Existing methods include regular expression-based or dictionary lookup-based methods that are not robust to dirty as well unseen data and are limited to a very less number of semantic data types to predict. Existing Machine Learning methods extract a large number of engineered features from data and build logistic regression, random forest or feedforward neural network for this purpose. In this paper, we introduce DCoM, a collection of multi-input NLP-based deep neural networks to detect semantic data types where instead of extracting a large number of features from the data, we feed the raw values of columns (or instances) to the model as texts. We train DCoM on 686,765 data columns extracted from the VizNet corpus with 78 different semantic data types. DCoM outperforms other contemporary results with a quite significant margin on the same dataset achieving a support-weighted F1 score of 0.925.

****

Efficient Wasserstein and Sinkhorn Policy Optimization

Jun Song,Chaoyue Zhao,Niao He

Trust-region methods based on Kullback-Leibler divergence are pervasively used to stabilize policy optimization in reinforcement learning. In this paper, we examine two natural extensions of policy optimziation with Wasserstein and Sinkhorn trust regions, namely Wasserstein policy optimization (WPO) and Sinkhorn policy optimization (SPO). Instead of restricting the policy to a parametric distribution class, we directly optimize the policy distribution and derive their close-form policy updates based on the Lagrangian duality. Theoretically, we show that WPO guarantees a monotonic performance improvement, and SPO provably converges to WPO as the entropic regularizer diminishes. Experiments across tabular domains and robotic locomotion tasks further demonstrate the performance improvement of both approaches, more robustness of WPO to sample insufficiency, and faster convergence of SPO, over state-of-art policy gradient methods.

****

Convergence of Generalized Belief Propagation Algorithm on Graphs with Motifs

Yitao Chen,Deepanshu Vasal

Belief propagation is a fundamental message-passing algorithm for numerous applications in machine learning. It is known that belief propagation algorithm is exact on tree graphs. However, belief propagation is run on loopy graphs in most applications. So, understanding the behavior of belief propagation on loopy graphs has been a major topic for researchers in different areas. In this paper, we study the convergence behavior of generalized belief propagation algorithm on graphs with motifs (triangles, loops, etc.) We show under a certain initialization, generalized belief propagation converges to the global optimum of the Bethe free energy for ferromagnetic Ising models on graphs with motifs.

****

iLQR-VAE : control-based learning of input-driven dynamics with applications to neural data

Marine Schimel,Ta-Chu Kao,Kristopher T Jensen,Guillaume Hennequin

Understanding how neural dynamics give rise to behaviour is one of the most fundamental questions in systems neuroscience. To achieve this, a common approach is to record neural populations in behaving animals, and model these data as emanating from a latent dynamical system whose state trajectories can then be related back to behavioural observations via some form of decoding. As recordings are typically performed in localized circuits that form only a part of the wider implicated network, it is important to simultaneously learn the local dynamics and infer any unobserved external input that might drive them. Here, we introduce iLQR-VAE, a novel control-based approach to variational inference in nonlinear dynamical systems, capable of learning both latent dynamics, initial conditions, and ongoing external inputs. As in recent deep learning approaches, our method is based on an input-driven sequential variational autoencoder (VAE). The main novelty lies in the use of the powerful iterative linear quadratic regulator algorithm (iLQR) in the recognition model. Optimization of the standard evidence lower-bound requires differentiating through iLQR solutions, which is made possible by recent advances in differentiable control. Importantly, having the recognition model be implicitly defined by the generative model greatly reduces the number of free parameters and allows for flexible, high-quality inference. This makes it possible for instance to evaluate the model on a single long trial after training on smaller chunks. We demonstrate the effectiveness of iLQR-VAE on a range of synthetic systems, with autonomous as well as input-driven dynamics. We further apply it to neural and behavioural recordings in non-human primates performing two different reaching tasks, and show that iLQR-VAE yields high-quality kinematic reconstructions from the neural data.

****

Likelihood Training of Schrödinger Bridge using Forward-Backward SDEs Theory

Tianrong Chen,Guan-Horng Liu,Evangelos Theodorou

Schrödinger Bridge (SB) is an entropy-regularized optimal transport problem that has received increasing attention in deep generative modeling for its mathematical flexibility compared to the Scored-based Generative Model (SGM). However, it remains unclear whether the optimization principle of SB relates to the modern training of deep generative models, which often rely on constructing log-likelihood objectives.This raises questions on the suitability of SB models as a principled alternative for generative applications. In this work, we present a novel computational framework for likelihood training of SB models grounded on Forward-Backward Stochastic Differential Equations Theory – a mathematical methodology appeared in stochastic optimal control that transforms the optimality condition of SB into a set of SDEs. Crucially, these SDEs can be used to construct the likelihood objectives for SB that, surprisingly, generalizes the ones for SGM as special cases. This leads to a new optimization principle that inherits the same SB optimality yet without losing applications of modern generative training techniques, and we show that the resulting training algorithm achieves comparable results on generating realistic images on MNIST, CelebA, and CIFAR10. Our code is available at https://github.com/ghliu/SB-FBSDE.

****

Imitation Learning from Observations under Transition Model Disparity

Tanmay Gangwani,Yuan Zhou,Jian Peng

Learning to perform tasks by leveraging a dataset of expert observations, also known as imitation learning from observations (ILO), is an important paradigm for learning skills without access to the expert reward function or the expert actions. We consider ILO in the setting where the expert and the learner agents operate in different environments, with the source of the discrepancy being the transition dynamics model. Recent methods for scalable ILO utilize adversarial learning to match the state-transition distributions of the expert and the learner, an approach that becomes challenging when the dynamics are dissimilar. In this work, we propose an algorithm that trains an intermediary policy in the learner environment and uses it as a surrogate expert for the learner. The intermediary policy is learned such that the state transitions generated by it are close to the state transitions in the expert dataset. To derive a practical and scalable algorithm, we employ concepts from prior work on estimating the support of a probability distribution. Experiments using MuJoCo locomotion tasks highlight that our method compares favorably to the baselines for ILO with transition dynamics mismatch.

****

Picking up the pieces: separately evaluating supernet training and architecture selection

Gabriel Meyer-Lee,Nick Cheney

Differentiable Neural Architecture Search (NAS) has emerged as a simple and efficient method for the automated design of neural networks. Recent research has demonstrated improvements on various aspects on the original algorithm (DARTS), but comparative evaluation of these advances remains costly and difficult. We frame supernet NAS as a two-stage search, decoupling the training of the supernet from the extraction of a final design from the supernet. We propose a set of metrics which utilize benchmark data sets to evaluate each stage of the search process independently. We demonstrate two metrics measuring separately the quality of the supernet's shared weights and the quality of the learned sampling distribution, as well as corresponding statistics approximating the reliance of the second stage search on these components of the supernet. These metrics facilitate both more robust evaluation of NAS algorithms and provide practical method for designing complete NAS algorithms from separate supernet training and architecture selection techniques.

****

Extreme normalization: approximating full-data batch normalization with single examples

Sergey Ioffe

While batch normalization has been successful in speeding up the training of neural networks, it is not well understood. We cast batch normalization as an approximation of the limiting case where the entire dataset is normalized jointly, and explore other ways to approximate the gradient from this limiting case. We demonstrate an approximation that removes the need to keep more than one example in memory at any given time, at the cost of a small factor increase in the training step computation, as well as a fully per-example training procedure, which removes the extra computation at the cost of a small drop in the final model accuracy. We further use our insights to improve batch renormalization for very small minibatches. Unlike previously proposed methods, our normalization does not change the function class of the inference model, and performs well in the absence of identity shortcuts.

****

Set Norm and Equivariant Skip Connections: Putting the Deep in Deep Sets

Lily H Zhang,Veronica Tozzo,John M. Higgins,Rajesh Ranganath

Permutation invariant neural networks are a promising tool for predictive modeling of set data. We show, however, that existing architectures struggle to perform well when they are deep. In this work, we address this issue for the two most widely used permutation invariant networks, Deep Sets and its transformer analogue Set Transformer. We take inspiration from previous efforts to scale neural network architectures by incorporating normalization layers and skip connections that work for sets. First, we motivate and develop set norm, a normalization tailored for sets. Then, we employ equivariant residual connections and introduce the ``clean path principle'' for their placement. With these changes, our many-layer Deep Sets++ and Set Transformer++ models reach comparable or better performance than their original counterparts on a diverse suite of tasks, from point cloud classification to regression on sets of images. We additionally introduce Flow-RBC, a new single-cell dataset and real-world application of permutation invariant prediction. On this task, our new models outperform existing methods as well as a clinical baseline. We open-source our data and code here: link-omitted-for-anonymity.

****

Fundamental Limits of Transfer Learning in Binary Classifications

Mohammadreza Mousavi Kalan,Salman Avestimehr,Mahdi Soltanolkotabi

A critical performance barrier in modern machine learning is scarcity of labeled data required for training state of the art massive models, especially in quickly emerging problems with lack of extensive data sets or scenarios where data collection and labeling is expensive/time consuming. Transfer learning is gaining traction as a promising technique to alleviate this barrier by utilizing the data of a related but different \emph{source} task to compensate for the lack of data in a \emph{target} task where there are few labeled training data. While there has been many recent algorithmic advances in this domain, a fundamental understanding of when and how much one can transfer knowledge from a related domain to reduce the amount of labeled training data is far from understood. We provide a precise answer to this question for binary classification problems by deriving a novel lower bound on the generalization error that can be achieved by \emph{any} transfer learning algorithm (regardless of its computational complexity) as a function of the amount of source and target samples. Our lower bound depends on a natural notion of distance that can be easily computed on real world data sets. Other key features of our lower bound are that it does not depend on the source/target data distributions and requires minimal assumptions that enables it application to a broad range of problems. We also consider a more general setting where there are more than one source domains for knowledge transfer to the target task and develop new bounds on generalization error in this setting. We also corroborate our theoretical findings on real image classification and action recognition data sets. These experiments demonstrate that our natural notion of distance is indicative of the difficulty of knowledge transfer between different pairs of source/target tasks, allowing us to investigate the effect of different sources on the target generalization error. Furthermore, to evaluate the sharpness of our bounds we compare our developed lower bounds with upper-bounds achieved by transfer learning base-lines that utilize weighted empirical risk minimization on the combination of source(s) and target data sets.

****

Targeted Environment Design from Offline Data

Izzeddin Gur,Ofir Nachum,Aleksandra Faust

In reinforcement learning (RL) the use of simulators is ubiquitous, allowing cheaper and safer agent training than training directly in the real target environment. However, this approach relies on the simulator being a sufficiently accurate reflection of the target environment, which is difficult to achieve in practice, resulting in the need to bridge sim2real gap. Accordingly, recent methods have proposed an alternative paradigm, utilizing offline datasets from the target environment to train an agent, avoiding online access to either the target or any simulated environment but leading to poor generalization outside the support of the offline data. We propose to combine the two paradigms: offline datasets and synthetic simulators, to reduce the sim2real gap by using limited offline data to train realistic simulators. We formalize our approach as offline targeted environment design(OTED), which automatically learns a distribution over simulator parameters to match a provided offline dataset, and then uses the learned simulator to train an RL agent in standard online fashion. We derive an objective for learning the simulator parameters which corresponds to minimizing a divergence between the target offline dataset and the state-action distribution induced by the simulator. We evaluate our method on standard offlineRL benchmarks and show that it learns using as few as 5 demonstrations, and yields up to 17 times higher score compared to strong existing offline RL, behavior cloning (BC), and domain randomization baseline, thus successfully leveraging both offline datasets and simulators for better RL

****

Randomized Signature Layers for Signal Extraction in Time Series Data

Enea Monzio Compagnoni,Luca Biggio,Antonio Orvieto,Thomas Hofmann,Josef Teichmann

Time series analysis is a widespread task in Natural Sciences, Social Sciences, and Engineering. A fundamental problem is finding an expressive yet efficient-to-compute representation of the input time series to use as a starting point to perform arbitrary downstream tasks.

In this paper, we build upon recent work using the signature of a path as a feature map and investigate a computationally efficient technique to approximate these features based on linear random projections. We present several theoretical results to justify our approach and empirically validate that our random projections can effectively retrieve the underlying signature of a path.

We show the surprising performance of the proposed random features on several tasks, including (1) mapping the controls of Stochastic Differential Equations to the corresponding solutions and (2) using the random signatures as time series representation for classification tasks. Besides providing a new tool to extract signatures and further validating the high level of expressiveness of such features, we believe our results provide interesting conceptual links between several existing research areas, suggesting new intriguing directions for future investigations.

****

Value Gradient weighted Model-Based Reinforcement Learning

Claas A Voelcker,Victor Liao,Animesh Garg,Amir-massoud Farahmand

Model-based reinforcement learning (MBRL) is a sample efficient technique to obtain control policies, yet unavoidable modeling errors often lead performance deterioration. The model in MBRL is often solely fitted to reconstruct dynamics, state observations in particular, while the impact of model error on the policy is not captured by the training objective. This leads to a mismatch between the intended goal of MBRL, enabling good policy and value learning, and the target of the loss function employed in practice, future state prediction. Naive intuition would suggest that value-aware model learning would fix this problem and, indeed, several solutions to this objective mismatch problem have been proposed based on theoretical analysis. However, they tend to be inferior in practice to commonly used maximum likelihood (MLE) based approaches. In this paper we propose the Value-gradient weighted Model Learning (VaGraM), a novel method for value-aware model learning which improves the performance of MBRL in challenging settings, such as small model capacity and the presence of distracting state dimensions. We analyze both MLE and value-aware approaches and demonstrate how they fail to account for exploration and the behavior of function approximation when learning value-aware models and highlight the additional goals that must be met to stabilize optimization in the deep learning setting. We verify our analysis by showing that our loss function is able to achieve high returns on the Mujoco benchmark suite while being more robust than maximum likelihood based approaches.

****

Expected Improvement-based Contextual Bandits

Hung Tran-The,Sunil Gupta,Santu Rana,Long Tran-Thanh,Svetha Venkatesh

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 when no assumption is made about the reward function other than it being bounded. With a linear reward function, we demonstrate that our algorithm achieves a near-optimal regret. In particular, our regret bound reduces a factor of $\sqrt{\text{log}(T)}$ compared to the popular OFUL algorithm \citep{Abbasi11} which uses the UCB approach, and reduces a factor of $\sqrt{d\text{log}(T)}$ compared to another popular algorithm \citep{agrawal13} which uses the TS approach. Here T is the horizon and d is the feature vector dimension. Further, when no assumptions are made about the form of reward, we use deep neural networks to model the reward function. We prove that this algorithm also achieves a near-optimal regret. Finally, we provide an empirical evaluation of the algorithms on both synthetic functions and various benchmark datasets. Our experiments show that our algorithms work well and consistently outperform existing approaches.

****

MCMC Should Mix: Learning Energy-Based Model with Neural Transport Latent Space MCMC

Erik Nijkamp,Ruiqi Gao,Pavel Sountsov,Srinivas Vasudevan,Bo Pang,Song-Chun Zhu,Ying Nian Wu

Learning energy-based model (EBM) requires MCMC sampling of the learned model as an inner loop of the learning algorithm. However, MCMC sampling of EBMs in high-dimensional data space is generally not mixing, because the energy function, which is usually parametrized by deep network, is highly multi-modal in the data space. This is a serious handicap for both theory and practice of EBMs. In this paper, we propose to learn EBM with a flow-based model (or in general latent variable model) serving as a backbone, so that the EBM is a correction or an exponential tilting of the flow-based model. We show that the model has a particularly simple form in the space of the latent variables of the generative model, and MCMC sampling of the EBM in the latent space mixes well and traverses modes in the data space. This enables proper sampling and learning of EBMs.

****

Generalizing Successor Features to continuous domains for Multi-task Learning

Melissa Mozifian,Dieter Fox,David Meger,Fabio Ramos,Animesh Garg

The deep reinforcement learning (RL) framework has shown great promise to tackle sequential decision-making problems, where the agent learns to behave optimally through interactions with the environment and receiving rewards.

The ability of an RL agent to learn different reward functions concurrently has many benefits, such as the decomposition of task rewards and skill reuse. One obstacle for achieving this, is the amount of data required as well as the capacit y of the model for solving multiple tasks. In this paper, we consider the problem of continuous control for various robot manipulation tasks with an explicit representation that promotes skill reuse while learning multiple tasks, related through the reward function. Our approach relies on two key concepts: successor features (SF), a value function representation that decouples the dynamics of the environment from the rewards, and an actor-critic framework that incorporates the learned SF representations.

We propose a practical implementation of successor features in continuous action spaces. We first show how to learn a decomposable representation required by SF. Our proposed methods, is able to learn decoupled state and reward features representations. We study this approach on a non-trivial continuous control problems with compositional structure built into the reward functions of various tasks.

****

Autonomous Learning of Object-Centric Abstractions for High-Level Planning

Steven James,Benjamin Rosman,George Konidaris

We propose a method for autonomously learning an object-centric representation of a continuous and high-dimensional environment that is suitable for planning. Such representations can immediately be transferred between tasks that share the same types of objects, resulting in agents that require fewer samples to learn a model of a new task. We first demonstrate our approach on a 2D crafting domain consisting of numerous objects where the agent learns a compact, lifted representation that generalises across objects. We then apply it to a series of Minecraft tasks to learn object-centric representations and object types - directly from pixel data - that can be leveraged to solve new tasks quickly. The resulting learned representations enable the use of a task-level planner, resulting in an agent capable of transferring learned representations to form complex, long-term plans.

****

Learning affective meanings that derives the social behavior using Bidirectional Encoder Representations from Transformers

Moeen Mostafavi,Michael D. Porter,Dawn T Robinson

Cultural sentiments of a society characterize social behaviors, but modeling sentiments to manifest every potential interaction remains an immense challenge. Affect Control Theory (ACT) offers a solution to this problem. ACT is a generative theory of culture and behavior based on a three-dimensional sentiment lexicon. Traditionally, the sentiments are quantified using survey data which is fed into a regression model to explain social behavior. The lexicons used in the survey are limited due to prohibitive cost. This paper uses a fine-tuned Bidirectional Encoder Representations from Transformers (BERT) model for developing a replacement for these surveys. This model achieves state-of-the-art accuracy in estimating affective meanings, expanding the affective lexicon, and allowing more behaviors to be explained.

****

Ensembles and Cocktails: Robust Finetuning for Natural Language Generation

John Hewitt,Xiang Lisa Li,Sang Michael Xie,Benjamin Newman,Percy Liang

When finetuning a pretrained language model for natural language generation tasks, one is currently faced with a tradeoff. Lightweight finetuning (e.g., prefix-tuning, adapters), which freezes all or most of the parameters of the pretrained model, has been shown to achieve stronger out-of-distribution (OOD) performance than full finetuning, which tunes all of the parameters. However, lightweight finetuning can underperform full finetuning in-distribution (ID). In this work, we present methods to combine the benefits of full and lightweight finetuning, achieving strong performance both ID and OOD. First, we show that an ensemble of the lightweight and full finetuning models achieves the best of both worlds: performance matching the better of full and lightweight finetuning, both ID and OOD. Second, we show that we can achieve similar improvements using a single model instead of two with our proposed cocktail finetuning, which augments full finetuning via distillation from a lightweight model. Finally, we provide some explanatory theory in a multiclass logistic regression setting with a large number of cl

asses, describing how distillation on ID data can transfer the OOD behavior of one model to another.
****

A fast and accurate splitting method for optimal transport: analysis and implementation
Vien V. Mai,Jacob Lindbäck,Mikael Johansson
We develop a fast and reliable method for solving large-scale optimal transport (OT) problems at an unprecedented combination of speed and accuracy. Built on the celebrated Douglas-Rachford splitting technique, our method tackles the original OT problem directly instead of solving an approximate regularized problem, as many state-of-the-art techniques do. This allows us to provide sparse transport plans and avoid numerical issues of methods that use entropic regularization. The algorithm has the same cost per iteration as the popular Sinkhorn method, and each iteration can be executed efficiently, in parallel. The proposed method enjoys an iteration complexity $O(1/\epsilon)$ compared to the best-known $O(1/\epsilon^2)$ of the Sinkhorn method. In addition, we establish a linear convergence rate for our formulation of the OT problem. We detail an efficient GPU implementation of the proposed method that maintains a primal-dual stopping criterion at no extra cost. Substantial experiments demonstrate the effectiveness of our method, both in terms of computation times and robustness.
****

The Geometry of Adversarial Subspaces
Dylan M. Paiton,David Schultheiss,Matthias Kuemmerer,Zac Cranko,Matthias Bethge
Artificial neural networks (ANNs) are constructed using well-understood mathematical operations, and yet their high-dimensional, non-linear, and compositional nature has hindered our ability to provide an intuitive description of how and why they produce any particular output. A striking example of this lack of understanding is our inability to design networks that are robust to adversarial input perturbations, which are often imperceptible to a human observer but cause significant undesirable changes in the network’s response. The primary contribution of this work is to further our understanding of the decision boundary geometry of ANN classifiers by utilizing such adversarial perturbations. For this purpose, we define adversarial subspaces, which are spanned by orthogonal directions of minimal perturbation to the decision boundary from any given input sample. We find that the decision boundary lies close to input samples in a large subspace, where the distance to the boundary grows smoothly and sub-linearly as one increases the dimensionality of the subspace. We undertake analysis to characterize the geometry of the boundary, which is more curved within the adversarial subspace than within a random subspace of equal dimensionality. To date, the most widely used defense against test-time adversarial attacks is adversarial training, where one incorporates adversarial attacks into the training procedure. Using our analysis, we provide new insight into the consequences of adversarial training by quantifying the increase in boundary distance within adversarial subspaces, the redistribution of proximal class labels, and the decrease in boundary curvature.
****

Implicit Bias of MSE Gradient Optimization in Underparameterized Neural Networks
Benjamin Bowman,Guido Montufar
We study the dynamics of a neural network in function space when optimizing the mean squared error via gradient flow. We show that in the underparameterized regime the network learns eigenfunctions of an integral operator T_K determined by the Neural Tangent Kernel at rates corresponding to their eigenvalues. For example, for uniformly distributed data on the sphere $S^{d - 1}$ and rotation invariant weight distributions, the eigenfunctions of T_K are the spherical harmonics. Our results can be understood as describing a spectral bias in the underparameterized regime. The proofs use the concept of ``Damped Deviations'' where deviations of the NTK matter less for eigendirections with large eigenvalues. Aside from the underparameterized regime, the damped deviations point-of-view allows us to extend certain results in the literature in the overparameterized setting.
****

Heterogeneous Wasserstein Discrepancy for Incomparable Distributions

Mokhtar Z. Alaya,Gilles Gasso,Maxime Berar,Alain Rakotomamonjy

Optimal Transport (OT) metrics allow for defining discrepancies between two probability measures. Wasserstein distance is for longer the celebrated OT-distance frequently-used in the literature, which seeks probability distributions to be supported on the $\text{\it same}$ metric space. Because of its high computational complexity, several approximate Wasserstein distances have been proposed based on entropy regularization or on slicing, and one-dimensional Wassserstein computation. In this paper, we propose a novel extension of Wasserstein distance to compare two incomparable distributions, that hinges on the idea of $\text{\it distributional slicing}$, embeddings, and on computing the closed-form Wassertein distance between the sliced distributions. We provide a theoretical analysis of this new divergence, called $\text{\it heterogeneous Wasserstein discrepancy (HWD)}$, and we show that it preserves several interesting properties including rotation-invariance. We show that the embeddings involved in HWD can be efficiently learned. Finally, we provide a large set of experiments illustrating the behavior of HWD as a divergence in the context of generative modeling and in query framework.

****

Understanding AdamW through Proximal Methods and Scale-Freeness

Zhenxun Zhuang,Mingrui Liu,Ashok Cutkosky,Francesco Orabona

Adam has been widely adopted for training deep neural networks due to less hyperparameter tuning and remarkable performance. To improve generalization, Adam is typically used in tandem with a squared ℓ_2 regularizer (referred to as Adam-ℓ_2). However, even better performance can be obtained with AdamW, which decouples the gradient of the regularizer from the update rule of Adam-ℓ_2. Yet, we are still lacking a complete explanation of the advantages of AdamW. In this paper, we tackle this question from both an optimization and an empirical point of view. First, we show how to re-interpret AdamW as an approximation of a proximal gradient method, which takes advantage of the closed-form proximal mapping of the regularizer instead of only utilizing its gradient information as in Adam-ℓ_2. Next, we consider the property of ``scale-freeness'' enjoyed by AdamW and by its proximal counterpart: their updates are invariant to component-wise rescaling of the gradients. We provide empirical evidence across a wide range of deep learning experiments showing a correlation between the problems in which AdamW exhibits an advantage over Adam-ℓ_2 and the degree to which we expect the gradients of the network to exhibit multiple scales, thus motivating the hypothesis that the advantage of AdamW could be due to the scale-free updates.

****

Discovering Latent Concepts Learned in BERT

Fahim Dalvi,Abdul Rafae Khan,Firoj Alam,Nadir Durrani,Jia Xu,Hassan Sajjad

A large number of studies that analyze deep neural network models and their ability to encode various linguistic and non-linguistic concepts provide an interpretation of the inner mechanics of these models. The scope of the analyses is limited to pre-defined concepts that reinforce the traditional linguistic knowledge and do not reflect on how novel concepts are learned by the model. We address this limitation by discovering and analyzing latent concepts learned in neural network models in an unsupervised fashion and provide interpretations from the model's perspective. In this work, we study: i) what latent concepts exist in the pre-trained BERT model, ii) how the discovered latent concepts align or diverge from classical linguistic hierarchy and iii) how the latent concepts evolve across layers.

Our findings show: i) a model learns novel concepts (e.g. animal categories and demographic groups), which do not strictly adhere to any pre-defined categorization (e.g. POS, semantic tags), ii) several latent concepts are based on multiple properties which may include semantics, syntax, and morphology, iii) the lower layers in the model dominate in learning shallow lexical concepts while the higher layers learn semantic relations and iv) the discovered latent concepts highlight potential biases learned in the model. We also release a novel BERT ConceptNet dataset consisting of 174 concept labels and 1M annotated instances.

****

Enforcing fairness in private federated learning via the modified method of differential multipliers

Borja Rodríguez Gálvez,Filip Granqvist,Rogier van Dalen,Matt Seigel

Federated learning with differential privacy, or private federated learning, provides a strategy to train machine learning models while respecting users' privacy. However, differential privacy can disproportionately degrade the performance of the models on under-represented groups, as these parts of the distribution are difficult to learn in the presence of noise. Existing approaches for enforcing fairness in machine learning models have considered the centralized setting, in which the algorithm has access to the users' data. This paper introduces an algorithm to enforce group fairness in private federated learning, where users' data does not leave their devices. First, the paper extends the modified method of differential multipliers to empirical risk minimization with fairness constraints, thus providing an algorithm to enforce fairness in the central setting. Then, this algorithm is extended to the private federated learning setting. The proposed algorithm, FPFL, is tested on a federated version of the Adult dataset and an "unfair" version of the FEMNIST dataset. The experiments on these datasets show how private federated learning accentuates unfairness in the trained models, and how FPFL is able to mitigate such unfairness.

****

FaceDet3D: Facial Expressions with 3D Geometric Detail Hallucination

ShahRukh Athar,Albert Pumarola,Francesc Moreno-noguer,Dimitris Samaras

Facial Expressions induce a variety of high-level details on the 3D face geometry. For example, a smile causes the wrinkling of cheeks or the formation of dimples, while being angry often causes wrinkling of the forehead. Morphable Models (3DMMs) of the human face fail to capture such fine details in their PCA-based representations and consequently cannot generate such details when used to edit expressions. In this work, we introduce FaceDet3D, a method that generates - from a single image - geometric facial details that are consistent with any desired target expression. The facial details are represented as a vertex displacement map and used then by a Neural Renderer to photo-realistically render novel images of any single image in any desired expression and view.

****

Picking Daisies in Private: Federated Learning from Small Datasets

Michael Kamp,Jonas Fischer,Jilles Vreeken

Federated learning allows multiple parties to collaboratively train a joint model without sharing local data. This enables applications of machine learning in settings of inherently distributed, undisclosable data such as in the medical domain. In practice, joint training is usually achieved by aggregating local models, for which local training objectives have to be in expectation similar to the joint (global) objective. Often, however, local datasets are so small that local objectives differ greatly from the global objective, resulting in federated learning to fail. We propose a novel approach that intertwines model aggregations with permutations of local models. The permutations expose each local model to a daisy chain of local datasets resulting in more efficient training in data-sparse domains. This enables training on extremely small local datasets, such as patient data across hospitals, while retaining the training efficiency and privacy benefits of federated learning.

****

A Broad Dataset is All You Need for One-Shot Object Detection

Claudio Michaelis,Matthias Bethge,Alexander S Ecker

Is it possible to detect arbitrary objects from a single example? A central problem of all existing attempts at one-shot object detection is the generalization gap: Object categories used during training are detected much more reliably than novel ones. We here show that this generalization gap can be nearly closed by increasing the number of object categories used during training. Doing so allows us to improve generalization from seen to unseen classes from 45% to 89% and improve the state-of-the-art on COCO by 5.4 AP50 (from 22.0 to 27.5).

We verify that the effect is caused by the number of categories and not the numb er of training samples, and that it holds for different models, backbones and datasets. This result suggests that the key to strong few-shot detection models may not lie in sophisticated metric learning approaches, but instead simply in scaling the number of categories. We hope that our findings will help to better understand the challenges of few-shot learning and encourage future data annotation efforts to focus on wider datasets with a broader set of categories rather than gathering more samples per category.

****

Generating Realistic 3D Molecules with an Equivariant Conditional Likelihood Model

James P. Roney,Paul Maragakis,Peter Skopp,David E. Shaw

The number of drug-like molecules that could potentially exist is thought to be above 10^{33}, precluding exhaustive computational or experimental screens for molecules with desirable pharmaceutical properties. Machine learning models that can propose novel molecules with specific characteristics are powerful new tools to break through the intractability of searching chemical space. Most of these models generate molecular graphs—representations that describe the topology of covalently bonded atoms in a molecule—because the bonding information in the graphs is required for many downstream applications, such as virtual screening and molecular dynamics simulation. These models, however, do not themselves generate 3D coordinates for the atoms within a molecule (which are also required for these applications), and thus they cannot easily incorporate information about 3D geometry when optimizing molecular properties. In this paper, we present GEN3D, a model that concurrently generates molecular graphs and 3D geometries, and is equivariant to rotations, translations, and atom permutations. The model extends a partially generated molecule by computing a conditional distribution over atom types, bonds, and spatial locations, and then sampling from that distribution to update the molecular graph and geometries, one atom at a time. We found that GEN3D proposes molecules that have much higher rates of chemical validity, and much better atom-distance distributions, than those generated with previous models. In addition, we validated our model’s geometric accuracy by forcing it to predict geometries for benchmark molecular graph inputs, and found that it also advances the state of the art on this test. We believe that the advantages that GEN3D provides over other models will enable it to contribute substantially to structure-based drug discovery efforts.

****

Generating Symbolic Reasoning Problems with Transformer GANs

Jens U. Kreber,Christopher Hahn

Constructing training data for symbolic reasoning domains is challenging: Existing instances are typically hand-crafted and too few to be trained on directly and synthetically generated instances are often hard to evaluate in terms of their meaningfulness. We study the capabilities of GANs and Wasserstein GANs equipped with Transformer encoders to generate sensible and challenging training data for symbolic reasoning domains. We conduct experiments on two problem domains where Transformers have been successfully applied recently: symbolic mathematics and temporal specifications in verification. Even without autoregression, our GAN models produce syntactically correct instances and we show that these can be used as meaningful substitutes for real training data when training a classifier. Using a GAN setting also allows us to alter the target distribution: We show that by adding a classifier uncertainty part to the generator objective, we obtain a dataset that is even harder to solve for a classifier than our original dataset.

****

The Role of Permutation Invariance in Linear Mode Connectivity of Neural Networks

Rahim Entezari,Hanie Sedghi,Olga Saukh,Behnam Neyshabur

In this paper, we conjecture that if the permutation invariance of neural networks is taken into account, SGD solutions will likely have no barrier in the linear interpolation between them. Although it is a bold conjecture, we show how extensive empirical attempts fall short of refuting it. We further provide a preliminary theoretical result to support our conjecture. Our conjecture has implicatio ns for the lottery ticket hypothesis, distributed training, and ensemble methods. The source code is available at \url{https://github.com/rahimentezari/PermutationInvariance}.

****

What Doesn't Kill You Makes You Robust(er): How to Adversarially Train against Data Poisoning

Jonas Geiping,Liam H Fowl,Gowthami Somepalli,Micah Goldblum,Michael Moeller,Tom Goldstein

Data poisoning is a threat model in which a malicious actor tampers with training data to manipulate outcomes at inference time. A variety of defenses against this threat model have been proposed, but each suffers from at least one of the following flaws: they are easily overcome by adaptive attacks, they severely reduce testing performance, or they cannot generalize to diverse data poisoning threat models. Adversarial training, and its variants, are currently considered the only empirically strong defense against (inference-time) adversarial attacks.

In this work, we extend the adversarial training framework to defend against (training-time) data poisoning. Our method desensitizes networks to the effects of such attacks by creating poisons during training and injecting them into training batches.

We show that this defense withstands adaptive attacks, generalizes to diverse threat models, and incurs a better performance trade-off than previous defenses.

****

Multi-Task Distribution Learning

Connor Shorten

Multi-Task Learning describes training on multiple tasks simultaneously to leverage the shared information between tasks. Tasks are typically defined as alternative ways to label data. Given an image of a face, a model could either classify the presence of sunglasses, or the presence of facial hair. This example highlights how the same input image can be posed as two separate binary classification problems. We present Multi-Task Distribution Learning, highlighting the similarities between Multi-Task Learning and preparing for Distribution Shift. Even with rapid advances in large-scale models, a Multi-Task Learner that is trained with object detection will outperform zero-shot inference on object detection. Similarly, we show how training with a data distribution aids with performance on that data distribution. We begin our experiments with a pairing of distribution tasks. We then show that this scales to optimizing 10 distribution tasks simultaneously. We further perform a task grouping analysis to see which augmentations train well together and which do not. Multi-Task Distribution Learning highlights the similarities between Distribution Shift and Zero-Shot task inference. These experiments will continue to improve with advances in generative modeling that enables simulating more interesting distribution shifts outside of standard augmentations. In addition, we discuss how the WILDS benchmark of Domain Generalizations and Subpopulation Shifts will aid in future work. Utilizing the prior knowledge of data augmentation and understanding multi-task interference is a promising direction to understand the phenomenon of Distribution Shift. To facilitate reproduction, we are open-sourcing code, leaderboards, and experimental data upon publication.

****

Offline Pre-trained Multi-Agent Decision Transformer

Linghui Meng,Muning Wen,Yaodong Yang,chenyang le,Xi yun Li,Haifeng Zhang,Ying Wen,Weinan Zhang,Jun Wang,Bo XU

Offline reinforcement learning leverages static datasets to learn optimal policies with no necessity to access the environment. This is desirable for multi-agent systems due to the expensiveness of agents' online interactions and the demand for sample numbers. Yet, in multi-agent reinforcement learning (MARL), the paradigm of offline pre-training with online fine-tuning has never been reported, nor datasets or benchmarks for offline MARL research are available. In this paper, we intend to investigate whether offline training is able to learn policy representations that elevate performance on downstream MARL tasks. We introduce the first offline dataset based on StarCraftII with diverse quality levels and propo se a multi-agent decision transformer (MADT) for effective offline learning. MADT integrates the powerful temporal representation learning ability of Transformer into both offline and online multi-agent learning, which promotes generalisation across agents and scenarios. The proposed method demonstrates superior performance than the state-of-the-art algorithms in offline MARL. Furthermore, when applied to online tasks, the pre-trained MADT largely improves sample efficiency, even in zero-shot task transfer. To our best knowledge, this is the first work to demonstrate the effectiveness of pre-trained models in terms of sample efficiency and generalisability enhancement in MARL.

****

Monotone deep Boltzmann machines
Zhili Feng,Ezra Winston,J Zico Kolter

Deep Boltzmann machines refer to deep multi-layered probabilistic models, governed by a pairwise energy function that describes the likelihood of all variables in the network. Due to the difficulty of inference in such systems, they have given way largely to \emph{restricted} deep Boltzmann machines (which do not permit intra-layer or skip connections). In this paper, we propose a class of model that allows for \emph{exact, efficient} mean-field inference and learning in \emph{general} deep Boltzmann machines. To do so, we use the tools of the recently proposed monotone Deep Equilibrium (DEQ) Model, an implicit-depth deep network that always guarantees the existence and uniqueness of its fixed points. We show that, for a class of general deep Boltzmann machine, the mean-field fixed point can be considered as the equivalent fixed point of a monotone DEQ, which gives us a recipe for deriving an efficient mean-field inference procedure with global convergence guarantees. In addition, we show that our procedure outperforms existing mean-field approximation methods while avoiding any issue of local optima. We apply this approach to simple deep convolutional Boltzmann architectures and demonstrate that it allows for tasks such as the joint completion and classification of images, all within a single deep probabilistic setting.

****

A Variance Reduction Method for Neural-based Divergence Estimation
Jeremiah Birrell,Markos A. Katsoulakis,Yannis Pantazis,Dipjyoti Paul,Anastasios Tsourtis

A central problem in machine learning is the computation of similarity or closeness between two (data) distributions. The applications span from generative modelling via adversarial training, representation learning, and robustness in out-of-distribution settings, to name a few. A palette of divergences, mutual information, and integral probability metrics are indispensable tools for measuring the ``distance'' between distributions and these are made tractable in high dimensional settings through variational representation formulas. Indeed, such formulas transform an estimation problem into an optimization problem. Unfortunately, the approximation of expectations that are inherent in variational formulas by statistical averages can be problematic due to high statistical variance, e.g., exponential for the Kullback-Leibler divergence and certain estimators. In this paper, we propose a new variance penalty term that acts directly on the variance of each component of the statistical estimator. The power of the variance penalty is controlled by a penalty coefficient which trades off bias and variance. We tested the proposed approach on several variational formulas and synthetic examples and showed that the overall error is decreased about an order of magnitude relative to the baseline statistical estimator. Impressive results are obtained for R\'enyi divergence with large order values due to the improved stability of the proposed estimator. Furthermore, in real biological datasets we are able to detect very rare sub-populations with a moderate sample size. Finally, we obtain improved (in terms of objective measures) disentangled representation of speech signals into text, speaker, and style components via variance-penalized mutual information minimization.

****

Fairness in Representation for Multilingual NLP: Insights from Controlled Experiments on Conditional Language Modeling
Ada Wan

We perform systematically and fairly controlled experiments with the 6-layer Transformer to investigate the hardness in conditional-language-modeling languages which have been traditionally considered morphologically rich (AR and RU) and poor (ZH). We evaluate through statistical comparisons across 30 possible language directions from the 6 languages of the United Nations Parallel Corpus across 5 data sizes on 3 representation levels --- character, byte, and word. Results show that performance is relative to the representation granularity of each of the languages, not to the language as a whole. On the character and byte levels, we are able to eliminate statistically significant performance disparity, hence demonstrating that a language cannot be intrinsically hard. The disparity that mirrors the morphological complexity hierarchy is shown to be a byproduct of word segmentation. Evidence from data statistics, along with the fact that word segmentation is qualitatively indeterminate, renders a decades-long debate on morphological complexity (unless it is being intentionally modeled in a word-based, meaning-driven context) irrelevant in the context of computing. The intent of our work is to help effect more objectivity and adequacy in evaluation as well as fairness and inclusivity in experimental setup in the area of language and computing so to uphold diversity in Machine Learning and Artificial Intelligence research. Multilinguality is real and relevant in computing not due to canonical, structural linguistic concepts such as morphology or "words" in our minds, but rather standards related to internationalization and localization, such as character encoding --- something which has thus far been sorely overlooked in our discourse and curricula.

****

When Complexity Is Good: Do We Need Recurrent Deep Learning For Time Series Outlier Detection?

Alexander Capstick,Samaneh Kouchaki,Mazdak Ghajari,David J. Sharp,Payam M. Barnaghi

Outlier detection is a critical part of understanding a dataset and extracting results. Outlier detection is used in different domains for various reasons; including detecting stolen credit cards, spikes of energy usage, web attacks, or in-home activity monitoring. Within this paper, we look at when it is appropriate to apply recurrent deep learning methods for time series outlier detection versus non-recurrent methods. Recurrent deep learning methods have a larger capacity for learning complex representations in time series data. We apply these methods to various synthetic and real-world datasets, including a dataset containing information about the in-home movement of people living with dementia in a clinical study cross-referenced with their recorded unplanned hospital admissions and infection episodes. We also introduce two new outlier detection methods, that can be useful in detecting contextual outliers in time series data where complex temporal relationships and local variations in the time series are important.

****

Deep convolutional recurrent neural network for short-interval EEG motor imagery classification

Ahmed Bahaa Selim,Ian van der Linde

In this paper, a high-performance short-interval motor imagery classifier is presented that has good potential for use in real-time EEG-based brain-computer interfaces (BCIs). A hybrid deep Convolutional Recurrent Neural Network with Temporal Attention (CRNN-TA) is described that achieves state-of-art performance in four-class classification (73% accuracy, 60% kappa, 3% higher than the winner of the BCI IV 2A competition). An adaptation of the guided grad-CAM method is proposed for decision visualization. A novel EEG data augmentation technique, shuffled-crossover, is introduced that leads to a 3% increase in classification accuracy (relative to a comparable baseline). Classification accuracies for different windows sizes and time intervals are evaluated. An attention mechanism is also proposed that could serve as a feedback loop during data capture for the rejection of bad trials (e.g., those in which participants were inattentive).

****

Node-Level Differentially Private Graph Neural Networks

Ameya Daigavane,Gagan Madan,Aditya Sinha,Abhradeep Guha Thakurta,Gaurav Aggarwal

,Prateek Jain

Graph neural networks (GNNs) are a popular technique for modelling graph-structured data that compute node-level predictions via aggregation of information from the local neighborhood of each node. However, this aggregation implies increased risk of revealing sensitive information, as a node can participate in the inference for multiple nodes. This implies that standard privacy preserving machine learning techniques like differentially private stochastic gradient descent (DP-SGD) – which are designed for situations where each node/data point participate in inference of one point only – either do not apply or lead to inaccurate solutions. In this work, we formally define the problem of learning 1-layer GNNs with node-level privacy, and provide a method for the problem with a strong differential privacy guarantee. Even though each node can be involved in the inference for multiple nodes, by employing a careful sensitivity analysis and a non-trivial extension of the privacy-by-amplification technique, our method is able to provide accurate solutions with solid privacy parameters. Empirical evaluation on standard benchmarks demonstrates that our method is indeed able to learn accurate privacy preserving GNNs, while still outperforming standard non-private methods that completely ignore graph information.

****

A composable autoencoder-based algorithm for accelerating numerical simulations

Rishikesh Ranade,Derek Christopher Hill,Haiyang He,Amir Maleki,Norman Chang,Jay Pathak

Numerical simulations for engineering applications solve partial differential equations (PDE) to model various physical processes. Traditional PDE solvers are very accurate but computationally costly. On the other hand, Machine Learning (ML) methods offer a significant computational speedup but face challenges with accuracy and generalization to different PDE conditions, such as geometry, boundary conditions, initial conditions and PDE source terms. In this work, we propose a novel ML-based approach, CoAE-MLSim, which is an unsupervised, lower-dimensional, local method, that is motivated from key ideas used in commercial PDE solvers. This allows our approach to learn better with relatively fewer samples of PDE solutions. The proposed ML-approach is compared against commercial solvers for better benchmarks as well as latest ML-approaches for solving PDEs. It is tested for a variety of complex engineering cases to demonstrate its computational speed, accuracy, scalability, and generalization across different PDE conditions. The results show that our approach captures physics accurately across all metrics of comparison (including measures such as results on section cuts and lines).

****

On the Practicality of Deterministic Epistemic Uncertainty

Janis Postels,Mattia Segu,TAO SUN,Luca Sieber,Luc Van Gool,Fisher Yu,Federico Tombari

A set of novel approaches for estimating epistemic uncertainty in deep neural networks with a single forward pass has recently emerged as a valid alternative to Bayesian Neural Networks. On the premise of informative representations, these deterministic uncertainty methods (DUMs) achieve strong performance on detecting out-of-distribution (OOD) data while adding negligible computational costs at inference time. However, it remains unclear whether DUMs are well calibrated and can seamlessly scale to real-world applications - both prerequisites for their practical deployment. To this end, we first provide a taxonomy of DUMs and evaluate their calibration under continuous distributional shifts. Then, we extend them to semantic segmentation. We find that, while DUMs scale to realistic vision tasks and perform well on OOD detection, the practicality of current methods is undermined by poor calibration under distributional shifts.

****

SERCNN: Stacked Embedding Recurrent Convolutional Neural Network in Depression Detection on Twitter

Heng Ee Tay,Mei Kuan Lim,Chun Yong Chong

Conventional approach of self-reporting-based screening for depression is not scalable, expensive, and requires one to be fully aware of their mental health. Motivated by previous studies that demonstrated great potentials for using social

media posts to monitor and predict one's mental health status, this study utilizes natural language processing and machine learning techniques on social media data to predict one's risk of depression. Most existing works utilize handcrafted features, and the adoption of deep learning in this domain is still lacking. Social media texts are often unstructured, ill-formed, and contain typos, making handcrafted features and conventional feature extraction methods inefficient. Moreover, prediction models built on these features often require a high number of posts per individual for accurate predictions. Therefore, this study proposes a Stacked Embedding Recurrent Convolutional Neural Network (SERCNN) for a more optimized prediction that has a better trade-off between the number of posts and accuracy. Feature vectors of two widely available pretrained embeddings trained on two distinct datasets are stacked, forming a meta-embedding vector that has a more robust and richer representation for any given word. We adapt Lai et al. (2015) RCNN approach that incorporates both the embedding vector and context learned from the neural network to form the final user representation before performing classification. We conducted our experiments on the Shen et al. (2017) depression Twitter dataset, the largest ground truth dataset used in this domain. Using SERCNN, our proposed model achieved a prediction accuracy of 78% when using only ten posts from each user, and the accuracy increases to 90% with an F1-measure of 0.89 when five hundred posts are analyzed.
****
Data Poisoning Won’t Save You From Facial Recognition
Evani Radiya-Dixit,Sanghyun Hong,Nicholas Carlini,Florian Tramer
Data poisoning has been proposed as a compelling defense against facial recognition models trained on Web-scraped pictures. Users can perturb images they post online, so that models will misclassify future (unperturbed) pictures.

We demonstrate that this strategy provides a false sense of security, as it ignores an inherent asymmetry between the parties: users' pictures are perturbed once and for all before being published (at which point they are scraped) and must thereafter fool all future models---including models trained adaptively against the users' past attacks, or models that use new technologies discovered after the attack.

We evaluate two systems for poisoning attacks against large-scale facial recognition, Fawkes (500,000+ downloads) and LowKey. We demonstrate how an "oblivious" model trainer can simply wait for future developments in computer vision to nullify the protection of pictures collected in the past. We further show that an adversary with black-box access to the attack can (i) train a robust model that resists the perturbations of collected pictures and (ii) detect poisoned pictures uploaded online.

We caution that facial recognition poisoning will not admit an "arms race" between attackers and defenders. Once perturbed pictures are scraped, the attack cannot be changed so any future successful defense irrevocably undermines users' privacy.
****
MetaMorph: Learning Universal Controllers with Transformers
Agrim Gupta,Linxi Fan,Surya Ganguli,Li Fei-Fei
Multiple domains like vision, natural language, and audio are witnessing tremendous progress by leveraging Transformers for large scale pre-training followed by task specific fine tuning. In contrast, in robotics we primarily train a single robot for a single task. However, modular robot systems now allow for the flexible combination of general-purpose building blocks into task optimized morphologies. However, given the exponentially large number of possible robot morphologies, training a controller for each new design is impractical. In this work, we propose MetaMorph, a Transformer based approach to learn a universal controller over a modular robot design space. MetaMorph is based on the insight that robot morphology is just another modality on which we can condition the output of a Transformer. Through extensive experiments we demonstrate that large scale pre-train

ing on a variety of robot morphologies results in policies with combinatorial generalization capabilities, including zero shot generalization to unseen robot morphologies. We further demonstrate that our pre-trained policy can be used for sample-efficient transfer to completely new robot morphologies and tasks.

****

Improving Long-Horizon Imitation Through Language Prediction

Donald Joseph Hejna III,Pieter Abbeel,Lerrel Pinto

Complex, long-horizon planning and its combinatorial nature pose steep challenges for learning-based agents. Difficulties in such settings are exacerbated in low data regimes where over-fitting stifles generalization and compounding errors hurt accuracy. In this work, we explore the use of an often unused source of auxiliary supervision: language. Inspired by recent advances in transformer-based models, we train agents with an instruction prediction loss that encourages learning temporally extended representations that operate at a high level of abstraction. Concretely, we demonstrate that instruction modeling significantly improves performance in planning environments when training with a limited number of demonstrations on the BabyAI and Crafter benchmarks. In further analysis we find that instruction modeling is most important for tasks that require complex reasoning, while understandably offering smaller gains in environments that require simple plans. Our benchmarks and code will be publicly released.

****

Causally Focused Convolutional Networks Through Minimal Human Guidance

Rimmon Saloman Bhosale,Mrinal Das

Convolutional Neural Networks (CNNs) are the state of the art in image classification mainly due to their ability to automatically extract features from the images and in turn, achieve accuracy higher than any method in history. However, the flip side is, they are correlational models which aggressively learn features that highly correlate with the labels. Such features may not be causally related to the labels as per human cognition. For example, in a subset of images, cows can be on grassland, but classifying an image as cow based on the presence of grassland is incorrect. To marginalize out the effect of all possible contextual features we need to gather a huge training dataset, which is not always possible. Moreover, this prohibits the model to justify the decision. This issue has some serious implications in certain domains such as medicine, where the amount of data can be limited but the model is expected to justify its decisions. In order to mitigate this issue, our proposal is to focus CNN to extract features that are causal from a human perspective. We propose a mechanism to accept guidance from humans in the form of activation masks to modify the learning process of CNN. The amount of additional guidance can be small and can be easily formed. Through detailed analysis, we show that this method not only improves the learning of causal features but also helps in learning efficiently with less data. We demonstrate the effectiveness of our method against multiple datasets using quantitative as well as qualitative results.

****

HYPOCRITE: Homoglyph Adversarial Examples for Natural Language Web Services in the Physical World

JINYONG KIM,JEONGHYEON KIM,MOSE GU,SANGHAK OHH,GILTEUN CHOI,JAEHOON JEONG

Recently, as Artificial Intelligence (AI) develops, many companies in various industries are trying to use AI by grafting it into their domains.

Also, for these companies, various cloud companies (e.g., Amazon, Google, IBM, and Microsoft) are providing AI services as the form of Machine-Learning-as-a-Service (MLaaS).

However, although these AI services are very advanced and well-made, security vulnerabilities such as adversarial examples still exist, which can interfere with normal AI services.

This paper demonstrates a HYPOCRITE for hypocrisy that generates homoglyph adversarial examples for natural language web services in the physical world. This hypocrisy can disrupt normal AI services provided by the cloud companies.

The key idea of HYPOCRITE is to replace English characters with other international characters that look similar to them in order to give the dataset noise to

the AI engines.
By using this key idea, parts of text can be appropriately replaced with subtext with malicious meaning through black-box attacks for natural language web services in order to cause misclassification.
In order to show attack potential by HYPOCRITE, this paper implemented a framework that makes homoglyph adversarial examples for natural language web services in the physical world and evaluated the performance under various conditions.
Through extensive experiments, it is shown that HYPOCRITE is more effective than other baseline in terms of both attack success rate and perturbed ratio.
****

Adaptive Behavior Cloning Regularization for Stable Offline-to-Online Reinforcement Learning
Yi Zhao,Rinu Boney,Alexander Ilin,Juho Kannala,Joni Pajarinen
Offline reinforcement learning, by learning from a fixed dataset, makes it possible to learn agent behaviors without interacting with the environment. However, depending on the quality of the offline dataset, such pre-trained agents may have limited performance and would further need to be fine-tuned online by interacting with the environment. During online fine-tuning, the performance of the pre-trained agent may collapse quickly due to the sudden distribution shift from offline to online data. While constraints enforced by offline RL methods such as a behaviour cloning loss prevent this to an extent, these constraints also significantly slow down online fine-tuning by forcing the agent to stay close to the behavior policy. We propose to adaptively weigh the behavior cloning loss during online fine-tuning based on the agent's performance and training stability. Moreover, we use a randomized ensemble of Q functions to further increase the sample efficiency of online fine-tuning by performing a large number of learning updates. Experiments show that the proposed method yields state-of-the-art offline-to-online reinforcement learning performance on the popular D4RL benchmark.
****

Continuous Control With Ensemble Deep Deterministic Policy Gradients
Piotr Januszewski,Mateusz Olko,Micha■ Królikowski,Jakub Swiatkowski,Marcin Andrychowicz,■ukasz Kuci■ski,Piotr Mi■o■
The growth of deep reinforcement learning (RL) has brought multiple exciting tools and methods to the field. This rapid expansion makes it important to understand the interplay between individual elements of the RL toolbox. We approach this task from an empirical perspective by conducting a study in the continuous control setting. We present multiple insights of fundamental nature, including: a commonly used additive action noise is not required for effective exploration and can even hinder training; the performance of policies trained using existing methods varies significantly across training runs, epochs of training, and evaluation runs; the critics' initialization plays the major role in ensemble-based actor-critic exploration, while the training is mostly invariant to the actors' initialization; a strategy based on posterior sampling explores better than the approximated UCB combined with the weighted Bellman backup; the weighted Bellman backup alone cannot replace the clipped double Q-Learning. As a conclusion, we show how existing tools can be brought together in a novel way, giving rise to the Ensemble Deep Deterministic Policy Gradients (ED2) method, to yield state-of-the-art results on continuous control tasks from \mbox{OpenAI Gym MuJoCo}. From the practical side, ED2 is conceptually straightforward, easy to code, and does not require knowledge outside of the existing RL toolbox.
****

HTLM: Hyper-Text Pre-Training and Prompting of Language Models
Armen Aghajanyan,Dmytro Okhonko,Mike Lewis,Mandar Joshi,Hu Xu,Gargi Ghosh,Luke Zettlemoyer
We introduce HTLM, a hyper-text language model trained on a large-scale web crawl. Modeling hyper-text has a number of advantages: (1) it is easily gathered at scale, (2) it provides rich document-level and end-task-adjacent supervision (e.g. 'class' and 'id' attributes often encode document category information), and (3) it allows for new structured prompting that follows the established semantics of HTML (e.g. to do zero-shot summarization by infilling '<title>' tags for a

webpage that contains the input text). We show that pretraining with a BART-style denoising loss directly on simplified HTML provides highly effective transfer for a wide range of end tasks and supervision levels. HTLM matches or exceeds the performance of comparably sized text-only LMs for zero-shot prompting and fine-tuning for classification benchmarks, while also setting new state-of-the-art performance levels for zero-shot summarization. We also find that hyper-text prompts provide more value to HTLM, in terms of data efficiency, than plain text prompts do for existing LMs, and that HTLM is highly effective at auto-prompting itself, by simply generating the most likely hyper-text formatting for any available training data. We will release all code and models to support future HTLM research.

****

Extending the WILDS Benchmark for Unsupervised Adaptation

Shiori Sagawa,Pang Wei Koh,Tony Lee,Irena Gao,Sang Michael Xie,Kendrick Shen,Ananya Kumar,Weihua Hu,Michihiro Yasunaga,Henrik Marklund,Sara Beery,Etienne David,Ian Stavness,Wei Guo,Jure Leskovec,Kate Saenko,Tatsunori Hashimoto,Sergey Levine,Chelsea Finn,Percy Liang

Machine learning systems deployed in the wild are often trained on a source distribution but deployed on a different target distribution. Unlabeled data can be a powerful point of leverage for mitigating these distribution shifts, as it is frequently much more available than labeled data and can often be obtained from distributions beyond the source distribution as well. However, existing distribution shift benchmarks with unlabeled data do not reflect the breadth of scenarios that arise in real-world applications. In this work, we present the WILDS 2.0 update, which extends 8 of the 10 datasets in the WILDS benchmark of distribution shifts to include curated unlabeled data that would be realistically obtainable in deployment. These datasets span a wide range of applications (from histology to wildlife conservation), tasks (classification, regression, and detection), and modalities (photos, satellite images, microscope slides, text, molecular graphs). The update maintains consistency with the original WILDS benchmark by using identical labeled training, validation, and test sets, as well as identical evaluation metrics. We systematically benchmark state-of-the-art methods that use unlabeled data, including domain-invariant, self-training, and self-supervised methods, and show that their success on WILDS is limited. To facilitate method development, we provide an open-source package that automates data loading and contains the model architectures and methods used in this paper. Code and leaderboards are available at https://wilds.stanford.edu.

****

Decomposing 3D Scenes into Objects via Unsupervised Volume Segmentation

Karl Stelzner,Kristian Kersting,Adam R. Kosiorek

We present ObSuRF, a method which turns a single image of a scene into a 3D model represented as a set of Neural Radiance Fields (NeRFs), with each NeRF corresponding to a different object. A single forward pass of an encoder network outputs a set of latent vectors describing the objects in the scene. These vectors are used independently to condition a NeRF decoder, defining the geometry and appearance of each object. We make learning more computationally efficient by deriving a novel loss, which allows training NeRFs on RGB-D inputs without explicit ray marching. After confirming that the model performs equal or better than state of the art on three 2D image segmentation benchmarks, we apply it to two multi-object 3D datasets: A multiview version of CLEVR, and a novel dataset in which scenes are populated by ShapeNet models. We find that after training ObSuRF on RGB-D views of training scenes, it is capable of not only recovering the 3D geometry of a scene depicted in a single input image, but also to segment it into objects, despite receiving no supervision in that regard.

****

Illiterate DALL-E Learns to Compose

Gautam Singh,Fei Deng,Sungjin Ahn

Although DALL-E has shown an impressive ability of composition-based systematic generalization in image generation, it requires the dataset of text-image pairs and the compositionality is provided by the text. In contrast, object-centric re presentation models like the Slot Attention model learn composable representations without the text prompt. However, unlike DALL-E, its ability to systematically generalize for zero-shot generation is significantly limited. In this paper, we propose a simple but novel slot-based autoencoding architecture, called SLATE, for combining the best of both worlds: learning object-centric representations that allow systematic generalization in zero-shot image generation without text. As such, this model can also be seen as an illiterate DALL-E model. Unlike the pixel-mixture decoders of existing object-centric representation models, we propose to use the Image GPT decoder conditioned on the slots for capturing complex interactions among the slots and pixels. In experiments, we show that this simple and easy-to-implement architecture not requiring a text prompt achieves significant improvement in in-distribution and out-of-distribution (zero-shot) image generation and qualitatively comparable or better slot-attention structure than the models based on mixture decoders.

****

Adaptive Learning of Tensor Network Structures

Meraj Hashemizadeh,Michelle Liu,Jacob Miller,Guillaume Rabusseau

Tensor Networks (TN) offer a powerful framework to efficiently represent very high-dimensional objects. TN have recently shown their potential for machine learning applications and offer a unifying view of common tensor decomposition models such as Tucker, tensor train (TT) and tensor ring (TR). However, identifying the best tensor network structure from data for a given task is challenging. In this work, we leverage the TN formalism to develop a generic and efficient adaptive algorithm to jointly learn the structure and the parameters of a TN from data. Our method is based on a simple greedy approach starting from a rank one tensor and successively identifying the most promising tensor network edges for small rank increments. Our algorithm can adaptively identify TN structures with small number of parameters that effectively optimize any differentiable objective function. Experiments on tensor decomposition, tensor completion and model compression tasks demonstrate the effectiveness of the proposed algorithm. In particular, our method outperforms the state-of-the-art evolutionary topology search [Li and Sun, 2020] for tensor decomposition of images (while being orders of magnitude faster) and finds efficient tensor network structures to compress neural networks outperforming popular TT based approaches [Novikov et al., 2015].

****

The Effects of Reward Misspecification: Mapping and Mitigating Misaligned Models

Alexander Pan,Kush Bhatia,Jacob Steinhardt

Reward hacking---where RL agents exploit gaps in misspecified proxy rewards---has been widely observed, but not yet systematically studied. To understand reward hacking, we construct four RL environments with different misspecified rewards. We investigate reward hacking as a function of agent capabilities: model capacity, action space resolution, and observation space noise. Typically, more capable agents are able to better exploit reward misspecifications, causing them to attain higher proxy reward and lower true reward. Moreover, we find instances of \emph{phase transitions}: capability thresholds at which the agent's behavior qualitatively shifts, leading to a sharp decrease in the true reward. Such phase transitions pose challenges to monitoring the safety of ML systems. To encourage further research on reward misspecification, address this, we propose an anomaly detection task for aberrant policies and offer several baseline detectors.

****

Optimizing Neural Networks with Gradient Lexicase Selection

Li Ding,Lee Spector

One potential drawback of using aggregated performance measurement in machine learning is that models may learn to accept higher errors on some training cases as compromises for lower errors on others, with the lower errors actually being instances of overfitting. This can lead both to stagnation at local optima and to poor generalization. Lexicase selection is an uncompromising method developed in evolutionary computation, which selects models on the basis of sequences of individual training case errors instead of using aggregated metrics such as loss and accuracy. In this paper, we investigate how the general idea of lexicase sele ction can fit into the context of deep learning to improve generalization. We propose Gradient Lexicase Selection, an optimization framework that combines gradient descent and lexicase selection in an evolutionary fashion. Experimental results show that the proposed method improves the generalization performance of various popular deep neural network architectures on three image classification benchmarks. Qualitative analysis also indicates that our method helps the networks learn more diverse representations.

****

Did I do that? Blame as a means to identify controlled effects in reinforcement learning

Oriol Corcoll Andreu,Youssef Sherif Mansour Mohamed,Raul Vicente

Identifying controllable aspects of the environment has proven to be an extraordinary intrinsic motivator to reinforcement learning agents. Despite repeatedly achieving State-of-the-Art results, this approach has only been studied as a proxy to a reward-based task and has not yet been evaluated on its own. We show that solutions relying on action-prediction fail to model critical controlled events. Humans, on the other hand, assign blame to their actions to decide what they controlled. This work proposes Controlled Effect Network (CEN), an unsupervised method based on counterfactual measures of blame to identify effects on the environment controlled by the agent. CEN is evaluated in a wide range of environments showing that it can accurately identify controlled effects. Moreover, we demonstrate CEN's capabilities as intrinsic motivator by integrating it in the state-of-the-art exploration method, achieving substantially better performance than action-prediction models.

****

Towards Coherent and Consistent Use of Entities in Narrative Generation

Pinelopi Papalampidi,Kris Cao,Tomáš Ko■iský

Large pre-trained language models (LMs) have demonstrated impressive capabilities in generating long, fluent text; however, there is little to no analysis on their ability to maintain entity coherence and consistency. In this work, we focus on the end task of narrative generation and systematically analyse the long-range entity coherence and consistency in generated stories. First, we propose a set of automatic metrics for measuring model performance in terms of entity usage. Given these metrics, we quantify the limitations of current LMs. Next, we propose augmenting a pre-trained LM with a dynamic entity memory in an end-to-end manner by using an auxiliary entity-related loss for guiding the reads and writes to the memory. We demonstrate that the dynamic entity memory increases entity coherence according to both automatic and human judgment and helps preserving entity-related information especially in settings with a limited context window. Finally, we also validate that our automatic metrics are correlated with human ratings and serve as a good indicator of the quality of generated stories.

****

Reinforcement Learning with Sparse Rewards using Guidance from Offline Demonstration

Desik Rengarajan,Gargi Vaidya,Akshay Sarvesh,Dileep Kalathil,Srinivas Shakkottai

A major challenge in real-world reinforcement learning (RL) is the sparsity of reward feedback. Often, what is available is an intuitive but sparse reward function that only indicates whether the task is completed partially or fully. However, the lack of carefully designed, fine grain feedback implies that most existing RL algorithms fail to learn an acceptable policy in a reasonable time frame. This is because of the large number of exploration actions that the policy has to perform before it gets any useful feedback that it can learn from. In this work, we address this challenging problem by developing an algorithm that exploits the offline demonstration data generated by {a sub-optimal behavior policy} for faster and efficient online RL in such sparse reward settings. The proposed algorithm, which we call the Learning Online with Guidance Offline (LOGO) algorithm, merges a policy improvement step with an additional policy guidance step by using the offline demonstration data. The key idea is that by obtaining guidance from - not imitating - the offline {data}, LOGO orients its policy in the manner of the sub-optimal {policy}, while yet being able to learn beyond and approa ch optimality. We provide a theoretical analysis of our algorithm, and provide a lower bound on the performance improvement in each learning episode. We also extend our algorithm to the even more challenging incomplete observation setting, where the demonstration data contains only a censored version of the true state observation. We demonstrate the superior performance of our algorithm over state-of-the-art approaches on a number of benchmark environments with sparse rewards {and censored state}. Further, we demonstrate the value of our approach via implementing LOGO on a mobile robot for trajectory tracking and obstacle avoidance, where it shows excellent performance.

****

Composing Features: Compositional Model Augmentation for Steerability of Music Transformers

Halley Young,Vincent Dumoulin,Pablo Samuel Castro,Jesse Engel,Cheng-Zhi Anna Huang

Music is a combinatorial art. Given a starting sequence, many continuations are possible, yet often only one is written down. With generative models, we can explore many. However, finding a continuation with specific combinations of features (such as rising pitches, with block chords played in syncopated rhythm) can take many trials.

To tackle the combinatorial nature of composing features, we propose a compositional approach to steering music transformers, building on lightweight fine-tuning methods such as prefix tuning and bias tuning. We introduce a novel contrastive loss function that enables us to steer compositional models over logical features using supervised learning. We examine the difficulty in steering based on whether features musically follow a prime or not, using existing music as a proxy. We show that with a relatively small number of extra parameters, our method allows bias tuning to perform successful fine-tuning in both the single-feature and compositional setting.

****

A Communication-Efficient Distributed Gradient Clipping Algorithm for Training Deep Neural Networks

Chunyang Liao,Zhenxun Zhuang,Mingrui Liu

In distributed training of deep neural networks or Federated Learning (FL), 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 FL setting is still in its infancy: it remains mysterious whether the gradient clipping scheme can take advantage of multiple machines to enjoy parallel speedup in the FL setting. The main technical difficulty lies at 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 $O\left(\frac{1}{N\epsilon^4}\right)$ iteration complexity for finding an ϵ-stationary point, where N is the number of machines. This indicates that our algorithm enjoys linear speedup. Our experiments on several benchmark datasets demonstrate that our algorithm indeed exhibits fast convergence speed in practice and validate our theory.

****

Why Should I Trust You, Bellman? Evaluating the Bellman Objective with Off-Policy Data

Scott Fujimoto,David Meger,Doina Precup,Ofir Nachum,Shixiang Shane Gu

In this work, we analyze the effectiveness of the Bellman equation as a proxy objective for value prediction accuracy in off-policy evaluation. While the Bellman equation is uniquely solved by the true value function over all state-action pairs, we show that in the finite data regime, the Bellman equation can be satisf ied exactly by infinitely many suboptimal solutions. This eliminates any guarantees relating Bellman error to the accuracy of the value function. We find this observation extends to practical settings; when computed over an off-policy dataset, the Bellman error bears little relationship to the accuracy of the value function. Consequently, we show that the Bellman error is a poor metric for comparing value functions, and therefore, an ineffective objective for off-policy evaluation. Finally, we discuss differences between Bellman error and the non-stationary objective used by iterative methods and deep reinforcement learning, and highlight how the effectiveness of this objective relies on generalization during training.

****

FedNAS: Federated Deep Learning via Neural Architecture Search

Chaoyang He,Erum Mushtaq,Jie Ding,Salman Avestimehr

Federated Learning (FL) is an effective learning framework used when data cannot be centralized due to privacy, communication costs, and regulatory restrictions. While there have been many algorithmic advances in FL, significantly less effort hasbeen made on model development, and most works in FL employ predefined model architectures discovered in the centralized environment. However, these predefinedarchitectures may not be the optimal choice for the FL setting since the user datadistribution at FL users is often non-identical and independent distribution (non-IID). This well-known challenge in FL has often been studied at the optimizationlayer. Instead, we advocate for a different (and complementary) approach. Wepropose Federated Neural Architecture Search (FedNAS) for automating the modeldesign process in FL. More specifically, FedNAS enables scattered workers tosearch for a better architecture in a collaborative fashion to achieve higher accuracy. Beyond automating and improving FL model design, FedNAS also provides anew paradigm for personalized FL via customizing not only the model weightsbut also the neural architecture of each user. As such, we also compare FedNASwith representative personalized FL methods, including perFedAvg (based on meta-learning), Ditto (bi-level optimization), and local fine-tuning. Our experiments ona non-IID dataset show that the architecture searched by FedNAS can outperformthe manually predefined architecture as well as existing personalized FL methods.To facilitate further research and real-world deployment, we also build a realisticdistributed training system for FedNAS, which will be publicly available andmaintained regularly.

****

1-bit LAMB: Communication Efficient Large-Scale Large-Batch Training with LAMB's Convergence Speed

Conglong Li,Ammar Ahmad Awan,Hanlin Tang,Samyam Rajbhandari,Yuxiong He

To train large models (like BERT and GPT-3) on hundreds of GPUs, communication has become a major bottleneck, especially on commodity systems with limited-bandwidth TCP network. On one side large batch-size optimization such as LAMB algorithm was proposed to reduce the frequency of communication. On the other side, communication compression algorithms such as 1-bit Adam help to reduce the volume of each communication. However, we find that simply using one of the techniques is not sufficient to solve the communication challenge, especially under low network bandwidth. Motivated by this we aim to combine the power of large-batch optimization and communication compression, but we find that existing compression strategies cannot be directly applied to LAMB due to its unique adaptive layerwise learning rates. To this end, we design a new communication-efficient algorithm, 1-bit LAMB, which introduces a novel way to support adaptive layerwise learning rates under compression. In addition, we introduce a new system implementation for compressed communication using the NCCL backend of PyTorch distributed, which improves both usability and performance. For BERT-Large pre-training task with batch sizes from 8K to 64K, our evaluations on up to 256 GPUs demonstrate that 1-bit LAMB with NCCL-based backend is able to achieve up to 4.6x communication volume reduction, up to 2.8x end-to-end time-wise speedup, and the same sample-wise convergence speed (and same fine-tuning task accuracy) compared to uncompressed LAMB.

****

One for Many: an Instagram inspired black-box adversarial attack

Alina Elena Baia,Alfredo Milani,Valentina Poggioni

It is well known that deep learning models are susceptible to adversarial attacks. To produce more robust and effective attacks, we propose a nested evolutionary algorithm able to produce multi-network (decision-based) black-box adversarial attacks based on Instagram inspired image filters. Due to the multi-network training, the system reaches a high transferability rate of attacks and, due to the composition of image filters, it is able to bypass standard detection mechanisms. Moreover, this kind of attack is semantically robust: our filter composition cannot be distinguished from any other filter composition used extensively every day to enhance images; this raises new security issues and challenges for real-world systems. Experimental results demonstrate that the method is also effective against
ensemble-adversarially trained models and it has a low cost in terms of queries to the victim model.

****

Linking Emergent and Natural Languages via Corpus Transfer

Shunyu Yao,Mo Yu,Yang Zhang,Karthik R Narasimhan,Joshua B. Tenenbaum,Chuang Gan

The study of language emergence aims to understand how human languages are shaped by perceptual grounding and communicative intent. Computational approaches to emergent communication (EC) predominantly consider referential games in limited domains and analyze the learned protocol within the game framework. As a result, it remains unclear how the emergent languages from these settings connect to natural languages or provide benefits in real-world language processing tasks, where statistical models trained on large text corpora dominate. In this work, we propose a novel way to establish such a link by corpus transfer, i.e. pretraining on a corpus of emergent language for downstream natural language tasks, which is in contrast to prior work that directly transfers speaker and listener parameters. Our approach showcases non-trivial transfer benefits for two different tasks – language modeling and image captioning. For example, in a low-resource setup (modeling 2 million natural language tokens), pre-training on an emergent language corpus with just 2 million tokens reduces model perplexity by 24.6% on average across ten natural languages. We also introduce a novel metric to predict the transferability of an emergent language by translating emergent messages to natural language captions grounded on the same images. We find that our translation-based metric highly correlates with the downstream performance on modeling natural languages (for instance $\rho = 0.83$ on Hebrew), while topographic similarity, a popular metric in previous works, shows surprisingly low correlation ($\rho = 0.003$), hinting that simple properties like attribute disentanglement from synthetic domains might not capture the full complexities of natural language. Our findings also indicate potential benefits of moving language emergence forward with natural language resources and models.

****

Offline Reinforcement Learning with Implicit Q-Learning

Ilya Kostrikov,Ashvin Nair,Sergey Levine

Offline reinforcement learning requires reconciling two conflicting aims: learning a policy that improves over the behavior policy that collected the dataset, while at the same time minimizing the deviation from the behavior policy so as to avoid errors due to distributional shift. This tradeoff is critical, because most current offline reinforcement learning methods need to query the value of unseen actions during training to improve the policy, and therefore need to either constrain these actions to be in-distribution, or else regularize their values. We propose a new offline RL method that never needs to evaluate actions outside of the dataset, but still enables the learned policy to improve substantially over the best behavior in the data through generalization. The main insight in our work is that, instead of evaluating unseen actions from the latest policy, we can approximate the policy improvement step implicitly by treating the state value function as a random variable, with randomness determined by the action (while still integrating over the dynamics to avoid excessive optimism), and then taking a state conditional upper expectile of this random variable to estimate the v alue of the best actions in that state. This leverages the generalization capacity of the function approximator to estimate the value of the best available action at a given state without ever directly querying a Q-function with this unseen action. Our algorithm alternates between fitting this upper expectile value function and backing it up into a Q-function, without any explicit policy. Then, we extract the policy via advantage-weighted behavioral cloning, which also avoids querying out-of-sample actions. We dub our method Implicit Q-learning (IQL). IQL is easy to implement, computationally efficient, and only requires fitting an additional critic with an asymmetric L2 loss. IQL demonstrates the state-of-the-art performance on D4RL, a standard benchmark for offline reinforcement learning. We also demonstrate that IQL achieves strong performance fine-tuning using online interaction after offline initialization.

****

Learning Distributionally Robust Models at Scale via Composite Optimization
Farzin Haddadpour,Mohammad Mahdi Kamani,Mehrdad Mahdavi,amin karbasi
To train machine learning models that are robust to distribution shifts in the data, distributionally robust optimization (DRO) has been proven very effective. However, the existing approaches to learning a distributionally robust model either require solving complex optimization problems such as semidefinite programming or a first-order method whose convergence scales linearly with the number of data samples-- which hinders their scalability to large datasets. In this paper, we show how different variants of DRO are simply instances of a finite-sum composite optimization for which we provide scalable methods. We also provide empirical results that demonstrate the effectiveness of our proposed algorithm with respect to the prior art in order to learn robust models from very large datasets.

****

FastRPB: a Scalable Relative Positional Encoding for Long Sequence Tasks
Maksim Zubkov,Daniil Gavrilov
Transformers achieve remarkable performance in various domains, including NLP, CV, audio processing, and graph analysis. However, they do not scale well on long sequence tasks due to their quadratic complexity w.r.t. the input’s length. Linear Transformers were proposed to address this limitation. However, these models have shown weaker performance on the long sequence tasks comparing to the original one. In this paper, we explore Linear Transformer models, rethinking their two core components. Firstly, we improved Linear Transformer with $\textbf{S}$hift-$\textbf{I}$nvariant $\textbf{K}$ernel $\textbf{F}$unction $\textbf{SIKF}$, which achieve higher accuracy without loss in speed. Secondly, we introduce $\textbf{FastRPB}$ which stands for $\textbf{Fast}$ $\textbf{R}$elative $\textbf{P}$ositional $\textbf{B}$ias, which efficiently adds positional information to self-attention using Fast Fourier Transformation. FastRPB is independent of the self-attention mechanism and can be combined with an original self-attention and all its efficient variants. FastRPB has $\mathcal{O}(N\log{N})$ computational complexity, requiring $\mathcal{O}(N)$ memory w.r.t. input sequence length N.

We compared introduced modifications with recent Linear Transformers in different settings: text classification, document retrieval, and image classification. Extensive experiments with FastRPB and SIKF demonstrate that our model significantly outperforms another efficient positional encodings method in accuracy, having up to x1.5 times higher speed and requiring up to x10 times less memory than the original Transformer.

****

STORM: Sketch Toward Online Risk Minimization
Gaurav Gupta,Benjamin Coleman,John Chen,Anshumali Shrivastava
Empirical risk minimization is perhaps the most influential idea in statistical learning, with applications to nearly all scientific and technical domains in the form of regression and classification models.

The growing concerns about the high energy cost of training and the increased prevalence of massive streaming datasets have led many ML practitioners to look for approximate ERM models that can achieve low cost on memory and latency for tra ining.

To this end, we propose STORM, an online sketching-based method for empirical risk minimization. STORM compresses a data stream into a tiny array of integer counters. This sketch is sufficient to estimate a variety of surrogate losses over the original dataset. We provide rigorous theoretical analysis and show that STORM can estimate a carefully chosen surrogate loss for regularized least-squares regression and a margin loss for classification.

We perform an exhaustive experimental comparison for regression and classification training on real-world datasets, achieving an approximate solution with a size even less than a data sample.

****

Counterfactual Plans under Distributional Ambiguity

Ngoc Bui,Duy Nguyen,Viet Anh Nguyen

Counterfactual explanations are attracting significant attention due to the flourishing applications of machine learning models in consequential domains. A counterfactual plan consists of multiple possibilities to modify a given instance so that the model's prediction will be altered. As the predictive model can be updated subject to the future arrival of new data, a counterfactual plan may become ineffective or infeasible, with respect to the future values of the model parameters. In this work, we study the counterfactual plans under model uncertainty, in which the distribution of the model parameters is partially prescribed using only the first- and second-moment information. First, we propose an uncertainty quantification tool to compute the lower and upper bounds of the probability of feasibility for any given counterfactual plan. We then provide corrective methods to adjust the counterfactual plan to improve the feasibility measure. The numerical experiments validate our bounds and demonstrate that our correction increases the robustness of the counterfactual plans in different real-world datasets.

****

Efficient representations for privacy-preserving inference

Han Xuanyuan,Francisco Vargas,Stephen Cummins

Deep neural networks have a wide range of applications across multiple domains such as computer vision and medicine. In many cases, the input of a model at inference time can consist of sensitive user data, which raises questions concerning the levels of privacy and trust guaranteed by such services. Much existing work has leveraged homomorphic encryption (HE) schemes that enable computation on encrypted data to achieve private inference for multi-layer perceptrons and CNNs. An early work along this direction was CryptoNets, which takes 250 seconds for one MNIST inference. The main limitation of such approaches is that of compute, which is due to the costly nature of the NTT (number theoretic transform) operations that constitute HE operations. Others have proposed the use of model pruning and efficient data representations to reduce the number of HE operations required. In this paper, we focus on improving upon existing work by proposing changes to the representations of intermediate tensors during CNN inference. We construct and evaluate private CNNs on the MNIST and CIFAR-10 datasets, and achieve over a two-fold reduction in the number of operations used for inferences of the CryptoNets architecture.

****

Delayed Geometric Discounts: An alternative criterion for Reinforcement Learning

Firas Jarboui,Ahmed Akakzia

The endeavor of artificial intelligence (AI) is to design autonomous agents capable of achieving complex tasks. Namely, reinforcement learning (RL) proposes a theoretical background to learn optimal behaviors. In practice, RL algorithms rely on geometric discounts to evaluate this optimality. Unfortunately, this does not cover decision processes where future returns are not exponentially less valuable.

Depending on the problem, this limitation induces sample-inefficiency (as feed-backs are exponentially decayed) and requires additional curricula/exploration mechanisms (to deal with sparse, deceptive or adversarial rewards).

In this paper, we tackle these issues by generalizing the discounted problem formulation with a family of delayed objective functions. We investigate the underl

ying RL problem to derive: 1) the optimal stationary solution and 2) an approximation of the optimal non-stationary control. The devised algorithms solved hard exploration problems on tabular environment and improved sample-efficiency on classic simulated robotics benchmarks.
****

Neural Parameter Allocation Search

Bryan A. Plummer,Nikoli Dryden,Julius Frost,Torsten Hoefler,Kate Saenko

Training neural networks requires increasing amounts of memory. Parameter sharing can reduce memory and communication costs, but existing methods assume networks have many identical layers and utilize hand-crafted sharing strategies that fail to generalize. We introduce Neural Parameter Allocation Search (NPAS), a novel task where the goal is to train a neural network given an arbitrary, fixed parameter budget. NPAS covers both low-budget regimes, which produce compact networks, as well as a novel high-budget regime, where additional capacity can be added to boost performance without increasing inference FLOPs. To address NPAS, we introduce Shapeshifter Networks (SSNs), which automatically learn where and how to share parameters in a network to support any parameter budget without requiring any changes to the architecture or loss function. NPAS and SSNs provide a complete framework for addressing generalized parameter sharing, and can also be combined with prior work for additional performance gains. We demonstrate the effectiveness of our approach using nine network architectures across four diverse tasks, including ImageNet classification and transformers.
****

Multi-Objective Model Selection for Time Series Forecasting

Oliver Borchert,David Salinas,Valentin Flunkert,Tim Januschowski,Stephan Günnemann

Research on time series forecasting has predominantly focused on developing methods that improve accuracy. However, other criteria such as training time or latency are critical in many real-world applications. We therefore address the question of how to choose an appropriate forecasting model for a given dataset among the plethora of available forecasting methods when accuracy is only one of many criteria. For this, our contributions are two-fold. First, we present a comprehensive benchmark, evaluating 7 classical and 6 deep learning forecasting methods on 44 heterogeneous, publicly available datasets. The benchmark code is open-sourced along with evaluations and forecasts for all methods. These evaluations enable us to answer open questions such as the amount of data required for deep learning models to outperform classical ones. Second, we leverage the benchmark evaluations to learn good defaults that consider multiple objectives such as accuracy and latency. By learning a mapping from forecasting models to performance metrics, we show that our method ParetoSelect is able to accurately select models from the Pareto front — alleviating the need to train or evaluate many forecasting models for model selection. To the best of our knowledge, ParetoSelect constitutes the first method to learn default models in a multi-objective setting.
****

TAMP-S2GCNets: Coupling Time-Aware Multipersistence Knowledge Representation with Spatio-Supra Graph Convolutional Networks for Time-Series Forecasting

Yuzhou Chen,Ignacio Segovia-Dominguez,Baris Coskunuzer,Yulia Gel

Graph Neural Networks (GNNs) are proven to be a powerful machinery for learning complex dependencies in multivariate spatio-temporal processes. However, most existing GNNs have inherently static architectures, and as a result, do not explicitly account for time dependencies of the encoded knowledge and are limited in their ability to simultaneously infer latent time-conditioned relations among entities. We postulate that such hidden time-conditioned properties may be captured by the tools of multipersistence, i.e, a emerging machinery in topological data analysis which allows us to quantify dynamics of the data shape along multiple geometric dimensions.

We make the first step toward integrating the two rising research directions, that is, time-aware deep learning and multipersistence, and propose a new model, Time-Aware Multipersistence Spatio-Supra Graph Convolutional Network (TAMP-S2GCNets). We summarize inherent time-conditioned topological properties of the data as time-aware multipersistence Euler-Poincar\'e surface and prove its stability. We then construct a supragraph convolution module which simultaneously accounts for the extracted intra- and inter- spatio-temporal dependencies in the data. Our extensive experiments on highway traffic flow, Ethereum token prices, and COVID-19 hospitalizations demonstrate that TAMP-S2GCNets outperforms the state-of-the-art tools in multivariate time series forecasting tasks.

****

Curriculum Learning: A Regularization Method for Efficient and Stable Billion-Scale GPT Model Pre-Training

Conglong Li,Minjia Zhang,Yuxiong He

Recent works have demonstrated great success in training high-capacity autoregressive language models (GPT, GPT-2, GPT-3) on a huge amount of unlabeled text corpus for text generation. Despite showing great results, autoregressive models are facing a growing training instability issue. Our study on GPT-2 models (117M and 1.5B parameters) show that larger model sizes, sequence lengths, batch sizes, and learning rates would lead to lower training stability and increasing divergence risks. To avoid divergence and achieve better generalization performance, one has to train with smaller batch sizes and learning rates, which leads to worse training efficiency and longer training time. To overcome this stability-efficiency dilemma, we present a study of a curriculum learning-based approach, which helps improves the pre-training convergence speed of autoregressive models. More importantly, we find that curriculum learning, as a regularization method, exerts a gradient variance reduction effect and enables to train autoregressive models with much larger batch sizes and learning rates without training instability, further improving the training speed. Our evaluations demonstrate that curriculum learning enables training GPT-2 models with 8x larger batch size and 4x larger learning rate, whereas the baseline approach struggles with training divergence. To achieve the same validation perplexity targets during pre-training, curriculum learning reduces the required number of tokens and wall clock time by up to 61% and 49%, respectively. To achieve the same or better zero-shot WikiText-103/LAMBADA evaluation results at the end of pre-training, curriculum learning reduces the required number of tokens and wall clock time by up to 54% and 70%, respectively.

****

Non-Linear Operator Approximations for Initial Value Problems

Gaurav Gupta,Xiongye Xiao,Radu Balan,Paul Bogdan

Time-evolution of partial differential equations is the key to model several dynamical processes, events forecasting but the operators associated with such problems are non-linear. We propose a Padé approximation based exponential neural operator scheme for efficiently learning the map between a given initial condition and activities at a later time. The multiwavelets bases are used for space discretization. By explicitly embedding the exponential operators in the model, we reduce the training parameters and make it more data-efficient which is essential in dealing with scarce real-world datasets. The Padé exponential operator uses a $\textit{recurrent structure with shared parameters}$ to model the non-linearity compared to recent neural operators that rely on using multiple linear operator layers in succession. We show theoretically that the gradients associated with the recurrent Padé network are bounded across the recurrent horizon. We perform experiments on non-linear systems such as Korteweg-de Vries (KdV) and Kuramoto–Sivashinsky (KS) equations to show that the proposed approach achieves the best performance and at the same time is data-efficient. We also show that urgent real-world problems like Epidemic forecasting (for example, COVID-19) can be formulated as a 2D time-varying operator problem. The proposed Padé exponential operators yield better prediction results ($\textbf{53\%} (\textbf{52\%})$ better MAE than best neural operator (non-neural operator deep learning model)) compared to state-of-the-art forecasting models.

****

A Generalised Inverse Reinforcement Learning Framework

Firas Jarboui,Vianney Perchet

The global objective of inverse Reinforcement Learning (IRL) is to estimate the

unknown cost function of some MDP based on observed trajectories generated by (approximate) optimal policies. The classical approach consists in tuning this cost function so that associated optimal trajectories (that minimise the cumulative discounted cost, i.e. the classical RL loss) are “similar” to the observed ones. Prior contributions focused on penalising degenerate solutions and improving algorithmic scalability. Quite orthogonally to them, we question the pertinence of characterising optimality with respect to the cumulative discounted cost as it induces an implicit bias against policies with longer mixing times. State of the art value based RL algorithms circumvent this issue by solving for the fixed point of the Bellman optimality operator, a stronger criterion that is not well defined for the inverse problem.
To alleviate this bias in IRL, we introduce an alternative training loss that puts more weights on future states which yields a reformulation of the (maximum entropy) IRL problem. The algorithms we devised exhibit enhanced performances (and similar tractability) than off-the-shelf ones in multiple OpenAI gym environments.

****

Learn Together, Stop Apart: a Novel Approach to Ensemble Pruning
Bulat Ibragimov,Gleb Gennadjevich Gusev
Gradient boosting is the most popular method of constructing ensembles that allow getting state-of-the-art results on many tasks. One of the critical parameters affecting the quality of the learned model is the number of models in the ensemble, or the number of boosting iterations. Unfortunately, the problem of selecting the optimal number of models still remains open and understudied. In this paper, we propose a new look at the hyperparameter selection problem in ensemble models. In contrast to the classical approaches that select the universal size of the ensemble from a hold-out validation subsample, our algorithm uses the hypothesis of heterogeneity of the sample space to adaptively set the required number of steps in one common ensemble for each group of objects individually. Experiments on popular implementations of gradient boosting show that the proposed method does not affect the complexity of learning algorithms and significantly increases quality on most standard benchmarks up to 1.5\%.

****

Proving Theorems using Incremental Learning and Hindsight Experience Replay
Eser Aygün,Laurent Orseau,Ankit Anand,Xavier Glorot,Vlad Firoiu,Lei M Zhang,Doina Precup,Shibl Mourad
Traditional automated theorem provers for first-order logic depend on speed-optimized search and many handcrafted heuristics that are designed to work best over a wide range of domains. Machine learning approaches in literature either depend on these traditional provers to bootstrap themselves or fall short on reaching comparable performance. In this paper, we propose a general incremental learning algorithm for training domain-specific provers for first-order logic without equality, based only on a basic given-clause algorithm, but using a learned clause-scoring function. Clauses are represented as graphs and presented to transformer networks with spectral features. To address the sparsity and the initial lack of training data as well as the lack of a natural curriculum, we adapt hindsight experience replay to theorem proving, so as to be able to learn even when no proof can be found. We show that provers trained this way can match and sometimes surpass state-of-the-art traditional provers on the TPTP dataset in terms of both quantity and quality of the proofs.

****

Constructing a Good Behavior Basis for Transfer using Generalized Policy Updates
Safa Alver,Doina Precup
We study the problem of learning a good set of policies, so that when combined together, they can solve a wide variety of unseen reinforcement learning tasks with no or very little new data. Specifically, we consider the framework of generalized policy evaluation and improvement, in which the rewards for all tasks of interest are assumed to be expressible as a linear combination of a fixed set of features. We show theoretically that, under certain assumptions, having access to a specific set of diverse policies, which we call a set of independent policie

s, can allow for instantaneously achieving high-level performance on all possible downstream tasks which are typically more complex than the ones on which the agent was trained. Based on this theoretical analysis, we propose a simple algorithm that iteratively constructs this set of policies. In addition to empirically validating our theoretical results, we compare our approach with recently proposed diverse policy set construction methods and show that, while others fail, our approach is able to build a behavior basis that enables instantaneous transfer to all possible downstream tasks. We also show empirically that having access to a set of independent policies can better bootstrap the learning process on downstream tasks where the new reward function cannot be described as a linear combination of the features. Finally, we demonstrate how this policy set can be useful in a lifelong reinforcement learning setting.

****

Collapse by Conditioning: Training Class-conditional GANs with Limited Data

Mohamad Shahbazi,Martin Danelljan,Danda Pani Paudel,Luc Van Gool

Class-conditioning offers a direct means to control a Generative Adversarial Network (GAN) based on a discrete input variable. While necessary in many applications, the additional information provided by the class labels could even be expected to benefit the training of the GAN itself. On the contrary, we observe that class-conditioning causes mode collapse in limited data settings, where unconditional learning leads to satisfactory generative ability. Motivated by this observation, we propose a training strategy for class-conditional GANs (cGANs) that effectively prevents the observed mode-collapse by leveraging unconditional learning. Our training strategy starts with an unconditional GAN and gradually injects the class conditioning into the generator and the objective function. The proposed method for training cGANs with limited data results not only in stable training but also in generating high-quality images, thanks to the early-stage exploitation of the shared information across classes. We analyze the observed mode collapse problem in comprehensive experiments on four datasets. Our approach demonstrates outstanding results compared with state-of-the-art methods and established baselines. The code is available at https://github.com/mshahbazi72/transitional-cGAN

****

FLAME-in-NeRF: Neural control of Radiance Fields for Free View Face Animation

ShahRukh Athar,Zhixin Shu,Dimitris Samaras

This paper presents a neural rendering method for controllable portrait video synthesis.Recent advances in volumetric neural rendering, such as neural radiance fields (NeRF), have enabled the photorealistic novel view synthesis of static scenes with impressive results. However, modeling dynamic and controllable objects as part of a scene with such scene representations is still challenging.

In this work, we design a system that enables 1) novel view synthesis for portrait video, of both the human subject and the scene they are in and 2) explicit control of the facial expressions through a low-dimensional expression representation.

We represent the distribution of human facial expressions using the expression parameters of a 3D Morphable Model (3DMMs) and condition the NeRF volumetric function on them.

Furthermore, we impose a spatial prior, brought by 3DMM fitting, to guide the network to learn disentangled control for static scene appearance and dynamic facial actions. We show the effectiveness of our method on free view synthesis of portrait videos with expression controls. To train a scene, our method only requires a short video of a subject captured by a mobile device.

****

SSFL: Tackling Label Deficiency in Federated Learning via Personalized Self-Supervision

Chaoyang He,Zhengyu Yang,Erum Mushtaq,Sunwoo Lee,Mahdi Soltanolkotabi,Salman Avestimehr

Federated Learning (FL) is transforming the ML training ecosystem from a centralized over-the-cloud setting to distributed training over edge devices in order to strengthen data privacy, reduce data migration costs, and break regulatory res trictions. An essential, but rarely studied, challenge in FL is label deficiency at the edge. This problem is even more pronounced in FL, compared to centralized training, due to the fact that FL users are often reluctant to label their private data and edge devices do not provide an ideal interface to assist with annotation. Addressing label deficiency is also further complicated in FL, due to the heterogeneous nature of the data at edge devices and the need for developing personalized models for each user. We propose a self-supervised and personalized federated learning framework, named SSFL, and a series of algorithms under this framework which work towards addressing these challenges. First, under the SSFL framework, we analyze the compatibility of various centralized self-supervised learning methods in FL setting and demonstrate that SimSiam networks performs the best with the standard FedAvg algorithm. Moreover, to address the data heterogeneity at the edge devices in this framework, we have innovated a series of algorithms that broaden existing supervised personalization algorithms into the setting of self-supervised learning including perFedAvg, Ditto, and local fine-tuning, among others. We further propose a novel personalized federated self-supervised learning algorithm, Per-SSFL, which balances personalization and consensus by carefully regulating the distance between the local and global representations of data. To provide a comprehensive comparative analysis of all proposed algorithms, we also develop a distributed training system and related evaluation protocol for SSFL. Using this training system, we conduct experiments on a synthetic non-I.I.D. dataset based on CIFAR-10, and an intrinsically non-I.I.D. dataset GLD-23K. Our findings show that the gap of evaluation accuracy between supervised learning and unsupervised learning in FL is both small and reasonable. The performance comparison indicates that representation regularization-based personalization method is able to outperform other variants. Ablation studies on SSFL are also conducted to understand the role of batch size, non-I.I.D.ness, and the evaluation protocol.

****

Public Data-Assisted Mirror Descent for Private Model Training
Ehsan Amid,Arun Ganesh,Rajiv Mathews,Swaroop Ramaswamy,Shuang Song,Thomas Steinke,Vinith Menon Suriyakumar,Om Thakkar,Abhradeep Guha Thakurta

In this paper, we revisit the problem of effectively using public data to improve the privacy/utility trade-offs for differentially private (DP) model training. Here, public data refers to auxiliary data sets that have no privacy concerns. We consider public training data sets that are from the *same distribution* as the private training data set.

For convex losses, we show that a variant of Mirror Descent provides population risk guarantees which are independent of the dimension of the model (p). Specifically, we apply Mirror Descent with the loss generated by the public data as the *mirror map*, and using DP gradients of the loss generated by the private (sensitive) data. To obtain dimension independence, we require $G_Q^2 \leq p$ public data samples, where G_Q is the Gaussian width of the smallest convex set Q such that the public loss functions are 1-strongly convex with respect to $\|\cdot\|_Q$. Our method is also applicable to non-convex losses, as it does not rely on convexity assumptions to ensure DP guarantees. We further show that our algorithm has a natural "noise stability" property: If in a bounded region around the current iterate, the public loss satisfies α_v-strong convexity in a direction v, then using noisy gradients instead of the exact gradients shifts our next iterate in the direction v by an amount proportional to $1/\alpha_v$ (in contrast with DP stochastic gradient descent (DP-SGD)), where the shift is isotropic). Analogous results in prior works had to explicitly learn the geometry using the public data in the form of preconditioner matrices.

We demonstrate the empirical efficacy of our algorithm by showing privacy/utility trade-offs on linear regression, and deep learning benchmark datasets (CIFAR-10, EMNIST, and WikiText-2). We show that our algorithm not only significantly improves over traditional DP-SGD, which does not have access to public data, but also improves over DP-SGD on models that have been pretrained with the public dat a to begin with.

****

E^{2}CM: Early Exit via Class Means for Efficient Supervised and Unsupervised Learning

Alperen Gormez,Erdem Koyuncu

State-of-the-art neural networks with early exit mechanisms often need considerable amount of training and fine-tuning to achieve good performance with low computational cost. We propose a novel early exit technique, E^{2}CM, based on the class means of samples. Unlike most existing schemes, E^{2}CM does not require gradient-based training of internal classifiers. This makes it particularly useful for neural network training in low-power devices, as in wireless edge networks. In particular, given a fixed training time budget, E^{2}CM achieves higher accuracy as compared to existing early exit mechanisms. Moreover, if there are no limitations on the training time budget, E^{2}CM can be combined with an existing early exit scheme to boost the latter's performance, achieving a better trade-off between computational cost and network accuracy. We also show that E^{2}CM can be used to decrease the computational cost in unsupervised learning tasks.

****

The MultiBERTs: BERT Reproductions for Robustness Analysis

Thibault Sellam,Steve Yadlowsky,Ian Tenney,Jason Wei,Naomi Saphra,Alexander D'Amour,Tal Linzen,Jasmijn Bastings,Iulia Raluca Turc,Jacob Eisenstein,Dipanjan Das, Ellie Pavlick

Experiments with pre-trained models such as BERT are often based on a single checkpoint. While the conclusions drawn apply to the artifact tested in the experiment (i.e., the particular instance of the model), it is not always clear whether they hold for the more general procedure which includes the architecture, training data, initialization scheme, and loss function. Recent work has shown that repeating the pre-training process can lead to substantially different performance, suggesting that an alternative strategy is needed to make principled statements about procedures. To enable researchers to draw more robust conclusions, we introduce MultiBERTs, a set of 25 BERT-Base checkpoints, trained with similar hyper-parameters as the original BERT model but differing in random weight initialization and shuffling of training data. We also define the Multi-Bootstrap, a non-parametric bootstrap method for statistical inference designed for settings where there are multiple pre-trained models and limited test data. To illustrate our approach, we present a case study of gender bias in coreference resolution, in which the Multi-Bootstrap lets us measure effects that may not be detected with a single checkpoint. The models and statistical library are available online, along with an additional set of 140 intermediate checkpoints captured during pre-training to facilitate research on learning dynamics.

****

Task-driven Discovery of Perceptual Schemas for Generalization in Reinforcement Learning

Wilka Torrico Carvalho,Andrew Kyle Lampinen,Kyriacos Nikiforou,Felix Hill,Murray Shanahan

Deep reinforcement learning (Deep RL) has recently seen significant progress in developing algorithms for generalization. However, most algorithms target a single type of generalization setting. In this work, we study generalization across three disparate task structures: (a) tasks composed of spatial and temporal compositions of regularly occurring object motions; (b) tasks composed of active perception of and navigation towards regularly occurring 3D objects; and (c) tasks composed of navigating through sequences of regularly occurring object-configurations. These diverse task structures all share an underlying idea of compositionality: task completion always involves combining reoccurring segments of task-oriented perception and behavior. We hypothesize that an agent can generalize within a task structure if it can discover representations that capture these reoccurring task-segments. For our tasks, this corresponds to representations for recognizing individual object motions, for navigation towards 3D objects, and for navigating through object-configurations. Taking inspiration from cognitive science, we term representations for reoccurring segments of an agent's experience, "p erceptual schemas". We propose Composable Perceptual Schemas (CPS), which learns a composable state representation where perceptual schemas are distributed across multiple, relatively small recurrent "subschema" modules. Our main technical novelty is an expressive attention function that enables subschemas to dynamically attend to features shared across all positions in the agent's observation. Our experiments indicate our feature-attention mechanism enables CPS to generalize better than recurrent architectures that attend to observations with spatial attention.

****

MaiT: integrating spatial locality into image transformers with attention masks

Ling Li,Ali Shafiee,Joseph H Hassoun

Though image transformers have shown competitive results with convolutional neural networks in computer vision tasks, lacking inductive biases such as locality still poses problems in terms of model efficiency especially for embedded applications. In this work, we address this issue by introducing attention masks to incorporate spatial locality into self-attention heads of transformers. Local dependencies are captured with masked attention heads along with global dependencies captured by original unmasked attention heads. With Masked attention image Transformer – MaiT, top-1 accuracy increases by up to 1.0\% compared to DeiT, without extra parameters, computation, or external training data. Moreover, attention masks regulate the training of attention maps, which facilitates the convergence and improves the accuracy of deeper transformers. Masked attention heads guide the model to focus on local information in early layers and promote diverse attention maps in latter layers. Deep MaiT improves the top-1 accuracy by up to 1.5\% compared to CaiT with fewer parameters and less FLOPs. Encoding locality with attention masks requires no extra parameter or structural change, and thus it can be combined with other techniques for further improvement in vision transformers.

****

High Probability Bounds for a Class of Nonconvex Algorithms with AdaGrad Stepsize

Ali Kavis,Kfir Yehuda Levy,Volkan Cevher

In this paper, we propose a new, simplified high probability analysis of AdaGrad for smooth, non-convex problems.

More specifically, we focus on a particular accelerated gradient (AGD) template (Lan, 2020), through which we recover the original AdaGrad and its variant with averaging, and prove a convergence rate of $\mathcal O (1/ \sqrt{T})$ with high probability without the knowledge of smoothness and variance.

We use a particular version of Freedman's concentration bound for martingale difference sequences (Kakade & Tewari, 2008) which enables us to achieve the best-known dependence of $\log (1 / \delta)$ on the probability margin δ.

We present our analysis in a modular way and obtain a complementary $\mathcal O (1 / T)$ convergence rate in the deterministic setting.

To the best of our knowledge, this is the first high probability result for AdaGrad with a truly adaptive scheme, i.e., completely oblivious to the knowledge of smoothness and uniform variance bound, which simultaneously has best-known dependence of $\log(1/ \delta)$.

We further prove noise adaptation property of AdaGrad under additional noise assumptions.

****

Map Induction: Compositional spatial submap learning for efficient exploration in novel environments

Sugandha Sharma,Aidan Curtis,Marta Kryven,Joshua B. Tenenbaum,Ila R Fiete

Humans are expert explorers and foragers. Understanding the computational cognitive mechanisms that support this capability can advance the study of the human mind and enable more efficient exploration algorithms. We hypothesize that humans explore new environments by inferring the structure of unobserved spaces through re-use of spatial information collected from previously explored spaces. Taking inspiration from the neuroscience of repeating map fragments and ideas about program induction, we present a novel ``Map Induction'' framework, which involves the generation of novel map proposals for unseen environments based on compositions of already-seen spaces in a Hierarchical Bayesian framework. The model thus explicitly reasons about unseen spaces through a distribution of strong spatial priors. We introduce a new behavioral Map Induction Task (MIT) that involves foraging for rewards to compare human performance with state-of-the-art existing models and Map Induction. We show that Map Induction better predicts human behavior than the non-inductive baselines. We also show that Map Induction, when used to augment state-of-the-art approximate planning algorithms, improves their performance.

****

How Did the Model Change? Efficiently Assessing Machine Learning API Shifts

Lingjiao Chen,Matei Zaharia,James Zou

ML prediction APIs from providers like Amazon and Google have made it simple to use ML in applications. A challenge for users is that such APIs continuously change over time as the providers update models, and changes can happen silently without users knowing. It is thus important to monitor when and how much the MLAPIs’ performance shifts. To provide detailed change assessment, we model MLAPI shifts as confusion matrix differences, and propose a principled algorithmic framework, MASA, to provably assess these shifts efficiently given a sample budget constraint.MASAemploys an upper-confidence bound based approach to adaptively determine on which data point to query the ML API to estimate shifts. Empirically, we observe significant ML API shifts from 2020 to 2021 among 12 out of 36 applications using commercial APIs from Google, Microsoft, Amazon, and other providers. These real-world shifts include both improvements and reductions in accuracy. Extensive experiments show that MASA can estimate such API shifts more accurately than standard approaches given the same budget

****

A Tale of Two Flows: Cooperative Learning of Langevin Flow and Normalizing Flow Toward Energy-Based Model

Jianwen Xie,Yaxuan Zhu,Jun Li,Ping Li

This paper studies the cooperative learning of two generative flow models, in which the two models are iteratively updated based on the jointly synthesized examples. The first flow model is a normalizing flow that transforms an initial simple density to a target density by applying a sequence of invertible transformations. The second flow model is a Langevin flow that runs finite steps of gradient-based MCMC toward an energy-based model. We start from proposing a generative framework that trains an energy-based model with a normalizing flow as an amortized sampler to initialize the MCMC chains of the energy-based model. In each learning iteration, we generate synthesized examples by using a normalizing flow initialization followed by a short-run Langevin flow revision toward the current energy-based model. Then we treat the synthesized examples as fair samples from the energy-based model and update the model parameters with the maximum likelihood learning gradient, while the normalizing flow directly learns from the synthesized examples by maximizing the tractable likelihood. Under the short-run non-mixing MCMC scenario, the estimation of the energy-based model is shown to follow the perturbation of maximum likelihood, and the short-run Langevin flow and the normalizing flow form a two-flow generator that we call CoopFlow. We provide an understating of the CoopFlow algorithm by information geometry and show that it is a valid generator as it converges to a moment matching estimator. We demonstrate that the trained CoopFlow is capable of synthesizing realistic images, reconstructing images, and interpolating between images.

****

FrugalMCT: Efficient Online ML API Selection for Multi-Label Classification Tasks

Lingjiao Chen,Matei Zaharia,James Zou

Multi-label classification tasks such as OCR and multi-object recognition are a major focus of the growing machine learning as a service industry. While many multi-label APIs are available, it is challenging for users to decide which API to use for their own data and budget, due to the heterogeneity in their prices and performance. Recent work has shown how to efficiently select and combine single label APIs to optimize performance and cost. However, its computation cost is exponential in the number of labels, and is not suitable for settings like OCR. In this work, we propose FrugalMCT, a principled framework that adaptively selects the APIs to use for different data in an online fashion while respecting the user’s budget. It allows combining ML APIs’ predictions for any single data point, and selects the best combination based on an accuracy estimator. We run systematic experiments using ML APIs from Google, Microsoft, Amazon, IBM, Tencent, and other providers for tasks including multi-label image classification, scene text recognition, and named entity recognition. Across these tasks, FrugalMCT can achieve over 90% cost reduction while matching the accuracy of the best single API, or up to 8% better accuracy while matching the best API’s cost.

****

Robust Learning Meets Generative Models: Can Proxy Distributions Improve Adversarial Robustness?

Vikash Sehwag,Saeed Mahloujifar,Tinashe Handina,Sihui Dai,Chong Xiang,Mung Chiang,Prateek Mittal

While additional training data improves the robustness of deep neural networks against adversarial examples, it presents the challenge of curating a large number of specific real-world samples. We circumvent this challenge by using additional data from proxy distributions learned by advanced generative models. We first seek to formally understand the transfer of robustness from classifiers trained on proxy distributions to the real data distribution. We prove that the difference between the robustness of a classifier on the two distributions is upper bounded by the conditional Wasserstein distance between them. Next we use proxy distributions to significantly improve the performance of adversarial training on five different datasets. For example, we improve robust accuracy by up to 7.5% and 6.7% in ℓ_{∞} and ℓ_2 threat model over baselines that are not using proxy distributions on the CIFAR-10 dataset. We also improve certified robust accuracy by 7.6% on the CIFAR-10 dataset. We further demonstrate that different generative models brings a disparate improvement in the performance in robust training. We propose a robust discrimination approach to characterize the impact and further provide a deeper understanding of why diffusion-based generative models are a better choice for proxy distribution than generative adversarial networks.

****

Semi-Empirical Objective Functions for Neural MCMC Proposal Optimization

Chris Cannella,Vahid Tarokh

Current objective functions used for training neural MCMC proposal distributions implicitly rely on architectural restrictions to yield sensible optimization results, which hampers the development of highly expressive neural MCMC proposal architectures. In this work, we introduce and demonstrate a semi-empirical procedure for determining approximate objective functions suitable for optimizing arbitrarily parameterized proposal distributions in MCMC methods. Our proposed Ab Initio objective functions consist of the weighted combination of functions following constraints on their global optima and transformation invariances that we argue should be upheld by general measures of MCMC efficiency for use in proposal optimization. Our experimental results demonstrate that Ab Initio objective functions maintain favorable performance and preferable optimization behavior compared to existing objective functions for neural MCMC optimization. We find that Ab Initio objective functions are sufficiently robust to enable the confident optimization of neural proposal distributions parameterized by deep generative networks extending beyond the regimes of traditional MCMC schemes.

****

Learning Higher-Order Dynamics in Video-Based Cardiac Measurement

Brian L. Hill,Xin Liu,Daniel McDuff

Computer vision methods typically optimize for first-order dynamics (e.g., optical flow). However, in many cases the properties of interest are subtle variations in higher-order changes, such as acceleration. This is true in the cardiac pulse, where the second derivative can be used as an indicator of blood pressure an d arterial disease. Recent developments in camera-based vital sign measurement have shown that cardiac measurements can be recovered with impressive accuracy from videos; however, the majority of research has focused on extracting summary statistics such as heart rate. Less emphasis has been put on the accuracy of waveform morphology that is necessary for many clinically impactful scenarios. In this work, we provide evidence that higher-order dynamics are better estimated by neural models when explicitly optimized for in the loss function. Furthermore, adding second-derivative inputs also improves performance when estimating second-order dynamics. By incorporating the second derivative of both the input frames and the target vital sign signals into the training procedure, our model is better able to estimate left ventricle ejection time intervals.

****

Chaos is a Ladder: A New Theoretical Understanding of Contrastive Learning via Augmentation Overlap

Yifei Wang,Qi Zhang,Yisen Wang,Jiansheng Yang,Zhouchen Lin

Recently, contrastive learning has risen to be a promising approach for large-scale self-supervised learning. However, theoretical understanding of how it works is still unclear. In this paper, we propose a new guarantee on the downstream performance without resorting to the conditional independence assumption that is widely adopted in previous work but hardly holds in practice. Our new theory hinges on the insight that the support of different intra-class samples will become more overlapped under aggressive data augmentations, thus simply aligning the positive samples (augmented views of the same sample) could make contrastive learning cluster intra-class samples together. Based on this augmentation overlap perspective, theoretically, we obtain asymptotically closed bounds for downstream performance under weaker assumptions, and empirically, we propose an unsupervised model selection metric ARC that aligns well with downstream accuracy. Our theory suggests an alternative understanding of contrastive learning: the role of aligning positive samples is more like a surrogate task than an ultimate goal, and the overlapped augmented views (i.e., the chaos) create a ladder for contrastive learning to gradually learn class-separated representations. The code for computing ARC is available at https://github.com/zhangq327/ARC.

****

Language-biased image classification: evaluation based on semantic representations

Yoann Lemesle,Masataka Sawayama,Guillermo Valle-Perez,Maxime Adolphe,Hélène Sauzéon,Pierre-Yves Oudeyer

Humans show language-biased image recognition for a word-embedded image, known as picture-word interference. Such interference depends on hierarchical semantic categories and reflects that human language processing highly interacts with visual processing. Similar to humans, recent artificial models jointly trained on texts and images, e.g., OpenAI CLIP, show language-biased image classification. Exploring whether the bias leads to interference similar to those observed in humans can contribute to understanding how much the model acquires hierarchical semantic representations from joint learning of language and vision. The present study introduces methodological tools from the cognitive science literature to assess the biases of artificial models. Specifically, we introduce a benchmark task to test whether words superimposed on images can distort the image classification across different category levels and, if it can, whether the perturbation is due to the shared semantic representation between language and vision. Our dataset is a set of word-embedded images and consists of a mixture of natural image datasets and hierarchical word labels with superordinate/basic category levels. Using this benchmark test, we evaluate the CLIP model. We show that presenting words distorts the image classification by the model across different category levels, but the effect does not depend on the semantic relationship between images and embedded words. This suggests that the semantic word representation in the CLIP visual processing is not shared with the image representation, although the word representation strongly dominates for word-embedded images.

****

Robbing the Fed: Directly Obtaining Private Data in Federated Learning with Mod ified Models

Liam H Fowl,Jonas Geiping,Wojciech Czaja,Micah Goldblum,Tom Goldstein

Federated learning has quickly gained popularity with its promises of increased user privacy and efficiency. Previous works have shown that federated gradient updates contain information that can be used to approximately recover user data in some situations. These previous attacks on user privacy have been limited in scope and do not scale to gradient updates aggregated over even a handful of data points, leaving some to conclude that data privacy is still intact for realistic training regimes. In this work, we introduce a new threat model based on minimal but malicious modifications of the shared model architecture which enable the server to directly obtain a verbatim copy of user data from gradient updates without solving difficult inverse problems. Even user data aggregated over large batches – where previous methods fail to extract meaningful content – can be reconstructed by these minimally modified models.

****

Reward Learning as Doubly Nonparametric Bandits: Optimal Design and Scaling Laws

Kush Bhatia,Wenshuo Guo,Jacob Steinhardt

Specifying reward functions for complex tasks like object manipulation or driving is challenging to do by hand. Reward learning seeks to address this by learning a reward model using human feedback on selected query policies. This shifts the burden of reward specification to the optimal design of the queries. We propose a theoretical framework for studying reward learning and the associated optimal experiment design problem. Our framework models rewards and policies as nonparametric functions belonging to subsets of Reproducing Kernel Hilbert Spaces (RKHSs). The learner receives (noisy) oracle access to a true reward and must output a policy that performs well under the true reward. For this setting, we first derive non-asymptotic excess risk bounds for a simple plug-in estimator based on ridge regression. We then solve the query design problem by optimizing these risk bounds with respect to the choice of query set and obtain a finite sample statistical rate, which depends primarily on the eigenvalue spectrum of a certain linear operator on the RKHSs. Despite the generality of these results, our bounds are stronger than previous bounds developed for more specialized problems. We specifically show that the well-studied problem of Gaussian process (GP) bandit optimization is a special case of our framework, and that our bounds either improve or are competitive with known regret guarantees for the Mat\'ern kernel.

****

Practical Conditional Neural Process Via Tractable Dependent Predictions

Stratis Markou,James Requeima,Wessel Bruinsma,Anna Vaughan,Richard E Turner

Conditional Neural Processes (CNPs; Garnelo et al., 2018a) are meta-learning models which leverage the flexibility of deep learning to produce well-calibrated predictions and naturally handle off-the-grid and missing data. CNPs scale to large datasets and train with ease. Due to these features, CNPs appear well-suited to tasks from environmental sciences or healthcare. Unfortunately, CNPs do not produce correlated predictions, making them fundamentally inappropriate for many estimation and decision making tasks. Predicting heat waves or floods, for example, requires modelling dependencies in temperature or precipitation over time and space. Existing approaches which model output dependencies, such as Neural Processes (NPs; Garnelo et al., 2018b) or the FullConvGNP (Bruinsma et al., 2021), are either complicated to train or prohibitively expensive. What is needed is an approach which provides dependent predictions, but is simple to train and computationally tractable. In this work, we present a new class of Neural Process models that make correlated predictions and support exact maximum likelihood training that is simple and scalable. We extend the proposed models by using invertible output transformations, to capture non-Gaussian output distributions. Our models can be used in downstream estimation tasks which require dependent function samples. By accounting for output dependencies, our models show improved predictive performance on a range of experiments with synthetic and real data.

****

PDAML: A Pseudo Domain Adaptation Paradigm for Subject-independent EEG-based Emotion Recognition
Yun Luo,Gengchen Wei,Bao-liang Lu

Domain adaptation (DA) and domain generalization (DG) methods have been successfully adopted to alleviate the domain shift problem caused by the subject variability of EEG signals in subject-independent affective brain-computer interfaces (aBCIs). Usually, the DA methods give relatively promising results than the DG methods but require additional computation resources each time a new subject comes. In this paper, we first propose a new paradigm called Pseudo Domain Adaptation (PDA), which is more suitable for subject-independent aBCIs. Then we propose the pseudo domain adaptation via meta-learning (PDAML) based on PDA. The PDAML consists of a feature extractor, a classifier, and a sum-decomposable structure called domain shift governor. We prove that a network with a sum-decomposable structure can compute the divergence between different domains effectively in theory. By taking advantage of the adversarial learning and meta-learning, the governor helps PDAML quickly generalize to a new domain using the target data through a few self-adaptation steps in the test phase. Experimental results on the public aBICs dataset demonstrate that our proposed method not only avoids the additional computation resources of the DA methods but also reaches a similar generalization performance of the state-of-the-art DA methods.

****

Reward Uncertainty for Exploration in Preference-based Reinforcement Learning
Xinran Liang,Katherine Shu,Kimin Lee,Pieter Abbeel

Conveying complex objectives to reinforcement learning (RL) agents often requires meticulous reward engineering. Preference-based RL methods are able to learn a more flexible reward model based on human preferences by actively incorporating human feedback, i.e. teacher's preferences between two clips of behaviors. However, poor feedback-efficiency still remains as a problem in current preference-based RL algorithms, as tailored human feedback is very expensive. To handle this issue, previous methods have mainly focused on improving query selection and policy initialization. At the same time, recent exploration methods have proven to be a recipe for improving sample-efficiency in RL. We present an exploration method specifically for preference-based RL algorithms. Our main idea is to design an intrinsic reward by measuring the novelty based on learned reward. Specifically, we utilize disagreement across ensemble of learned reward models. Our intuition is that disagreement in learned reward model reflects uncertainty in tailored human feedback and could be useful for exploration. Our experiments show that reward uncertainty exploration improves both feedback- and sample-efficiency of preference-based RL algorithms on complex robot manipulation tasks from Meta-World benchmarks, compared with other existing exploration methods that measure the novelty of state visitation.

****

Resolving label uncertainty with implicit generative models
Esther Rolf,Nikolay Malkin,Alexandros Graikos,Ana Jojic,Caleb Robinson,Nebojsa Jojic

In prediction problems, coarse and imprecise sources of input can provide rich information about labels, but are not readily used by discriminative learners. In this work, we propose a method for jointly inferring labels across a collection of data samples, where each sample consists of an observation and a prior belief about the label. By implicitly assuming the existence of a generative model for which a differentiable predictor is the posterior, we derive a training objective that allows learning under weak beliefs. This formulation unifies various machine learning settings: the weak beliefs can come in the form of noisy or incomplete labels, likelihoods given by a different prediction mechanism on auxiliary input, or common-sense priors reflecting knowledge about the structure of the problem at hand. We demonstrate the proposed algorithms on diverse problems: classification with negative training examples, learning from rankings, weakly and self-supervised aerial imagery segmentation, co-segmentation of video frames, and coarsely supervised text classification.

****

Congested bandits: Optimal routing via short-term resets
Pranjal Awasthi,Kush Bhatia,Sreenivas Gollapudi,Kostas Kollias

For traffic routing platforms, the choice of which route to recommend to a user depends on the congestion on these routes -- indeed, an individual's utility depends on the number of people using the recommended route at that instance. Motivated by this, we introduce the problem of Congested Bandits where each arm's reward is allowed to depend on the number of times it was played in the past Δ timesteps. This dependence on past history of actions leads to a dynamical system where an algorithm's present choices also affect its future pay-offs, and requires an algorithm to plan for this. We study the congestion aware formulation in the multi-armed bandit (MAB) setup and in the contextual bandit setup with linear rewards. For the multi-armed setup, we propose a UCB style algorithm and show that its policy regret scales as $\tilde{O}(\sqrt{K \Delta T})$. For the linear contextual bandit setup, our algorithm, based on an iterative least squares planner, achieves policy regret $\tilde{O}(\sqrt{dT} + \Delta)$. From an experimental standpoint, we corroborate the no-regret properties of our algorithms via a simulation study.

****

Tackling Oversmoothing of GNNs with Contrastive Learning
Lecheng Zheng,Dongqi Fu,Jingrui He

Graph neural networks (GNNs) integrate the comprehensive relation of graph data and the representation learning capability of neural networks, which is one of the most popular deep learning methods and achieves state-of-the-art performance in many applications, such as natural language processing and computer vision. In real-world scenarios, increasing the depth (i.e., the number of layers) of GNNs is sometimes necessary to capture more latent knowledge of the input data to mitigate the uncertainty caused by missing values.

However, involving more complex structures and more parameters will decrease the performance of GNN models. One reason called oversmoothing is recently proposed, whose research still remains nascent. In general, oversmoothing makes the final representations of nodes indiscriminative to hurt the node classification and link prediction performance.

In this paper, we first survey the current de-oversmoothing methods and propose three major metrics to evaluate a de-oversmoothing method, i.e., constant divergence indicator, easy-to-determine divergence indicator, and model-agnostic strategy. Then, we propose the Topology-guided Graph Contrastive Layer, named TGCL, which is the first de-oversmoothing method maintaining the three mentioned metrics. With the contrastive learning manner, we provide the theoretical analysis of the effectiveness of the proposed method. Last but not least, we design extensive experiments to illustrate the empirical performance of TGCL comparing with state-of-the-art baselines.

****

AriEL: volume coding for sentence generation comparisons
Luca Celotti,Simon Brodeur,Jean Rouat

Saving sequences of data to a point in a continuous space makes it difficult to retrieve them via random sampling. Mapping the input to a volume makes it easier, which is the strategy followed by Variational Autoencoders. However optimizing for prediction and for smoothness, forces them to trade-off between the two. We analyze the ability of standard deep learning techniques to generate sentences through latent space sampling. We compare toAriEL, an entropic coding method to construct volumes without the need for extra loss terms. We benchmark on a toy grammar, to automatically evaluate the language learned and generated, and find where it is stored in the latent space. Then, we benchmark on a dataset of human dialogues and using GPT-2 inside AriEL. Our results indicate that the random access to stored information can be improved since AriEL is able to generate a wider variety of correct language by randomly sampling the latent space. This supports the hypothesis that encoding information into volumes, leads to improved retrieval of learned information with random sampling.

****

Dense Gaussian Processes for Few-Shot Segmentation

Joakim Johnander,Johan Edstedt,Michael Felsberg,Fahad Khan,Martin Danelljan

Few-shot segmentation is a challenging dense prediction task, which entails segmenting a novel query image given only a small annotated support set. The key problem is thus to design a method that aggregates detailed information from the support set, while being robust to large variations in appearance and context. To this end, we propose a few-shot segmentation method based on dense Gaussian process (GP) regression. Given the support set, our dense GP learns the mapping from local deep image features to mask values, capable of capturing complex appearance distributions. Furthermore, it provides a principled means of capturing uncertainty, which serves as another powerful cue for the final segmentation, obtained by a CNN decoder. Instead of a one-dimensional mask output, we further exploit the end-to-end learning capabilities of our approach to learn a high-dimensional output space for the GP. Our approach sets a new state-of-the-art for both 1-shot and 5-shot FSS on the PASCAL-5^i and COCO-20^i benchmarks, achieving an absolute gain of $+14.9$ mIoU in the COCO-20^i 5-shot setting. Furthermore, the segmentation quality of our approach scales gracefully when increasing the support set size, while achieving robust cross-dataset transfer.

****

SoftHebb: Bayesian inference in unsupervised Hebbian soft winner-take-all networks

Timoleon Moraitis,Dmitry Toichkin,Yansong Chua,Qinghai Guo

State-of-the-art artificial neural networks (ANNs) require labelled data or feedback between layers, are often biologically implausible, and are vulnerable to adversarial attacks that humans are not susceptible to. On the other hand, Hebbian learning in winner-take-all (WTA) networks, is unsupervised, feed-forward, and biologically plausible. However, a modern objective optimization theory for WTA networks has been missing, except under very limiting assumptions. Here we derive formally such a theory, based on biologically plausible but generic ANN elements. Through Hebbian learning, network parameters maintain a Bayesian generative model of the data. There is no supervisory loss function, but the network does minimize cross-entropy between its activations and the input distribution. The key is a "soft" WTA where there is no absolute "hard" winner neuron, and a specific type of Hebbian-like plasticity of weights and biases. We confirm our theory in practice, where, in handwritten digit (MNIST) recognition, our Hebbian algorithm, SoftHebb, minimizes cross-entropy without having access to it, and outperforms the more frequently used, hard-WTA-based method. Strikingly, it even outperforms supervised end-to-end backpropagation, under certain conditions. Specifically, in a two-layered network, SoftHebb outperforms backpropagation when the training dataset is only presented once, when the testing data is noisy, and under gradient-based adversarial attacks. Notably, adversarial attacks that confuse SoftHebb are also confusing to the human eye. Finally, the model can generate interpolations of objects from its input distribution. All in all, SoftHebb extends Hebbian WTA theory with modern machine learning tools, thus making these networks relevant to pertinent issues in deep learning.

****

Decentralized Learning for Overparameterized Problems: A Multi-Agent Kernel Approximation Approach

Prashant Khanduri,Haibo Yang,Mingyi Hong,Jia Liu,Hoi To Wai,Sijia Liu

This work develops a novel framework for communication-efficient distributed learning where the models to be learned are overparameterized. We focus on a class of kernel learning problems (which includes the popular neural tangent kernel (NTK) learning as a special case) and propose a novel {\it multi-agent kernel approximation} technique that allows the agents to distributedly estimate the full kernel function, and subsequently perform decentralized optimization, without directly exchanging any local data or parameters. The proposed framework is a significant departure from the classical consensus-based approaches, because the agents do not exchange problem parameters, and no consensus is required. We analyze the optimization and the generalization performance of the proposed framework for the ℓ_2 loss. We show that with M agents and N total samples when certain generalized inner-product kernels (resp. the random features kernel) are us ed, each agent needs to communicate $\mathcal{O}\big({N^2}/{M}\big)$ bits (resp. $\mathcal{O}\big(N \sqrt{N}/M \big)$ real values) to achieve minimax optimal generalization performance. We validate the theoretical results on 90 UCI benchmarking datasets (with average data size $N \approx 1000$) and show that each agent needs to share a total of $200N/M$ bits (resp. $3N/M$ real values) to closely match the performance of the centralized algorithms, and these numbers are independent of parameter and feature dimensions.

****

Asymmetry Learning for Counterfactually-invariant Classification in OOD Tasks
S Chandra Mouli,Bruno Ribeiro

Generalizing from observed to new related environments (out-of-distribution) is central to the reliability of classifiers. However, most classifiers fail to predict label Y from input X when the change in environment is due a (stochastic) input transformation $T^\text{te} \circ X'$ not observed in training, as in training we observe $T^\text{tr} \circ X'$, where X' is a hidden variable. This work argues that when the transformations in train T^tr and test T^te are (arbitrary) symmetry transformations induced by a collection of known m equivalence relations, the task of finding a robust OOD classifier can be defined as finding the simplest causal model that defines a causal connection between the target labels and the symmetry transformations that are associated with label changes. We then propose a new learning paradigm, asymmetry learning, that identifies which symmetries the classifier must break in order to correctly predict Y in both train and test. Asymmetry learning performs a causal model search that, under certain identifiability conditions, finds classifiers that perform equally well in-distribution and out-of-distribution. Finally, we show how to learn counterfactually-invariant representations with asymmetry learning in two physics tasks.

****

Permutation-Based SGD: Is Random Optimal?
Shashank Rajput,Kangwook Lee,Dimitris Papailiopoulos

A recent line of ground-breaking results for permutation-based SGD has corroborated a widely observed phenomenon: random permutations offer faster convergence than with-replacement sampling. However, is random optimal? We show that this depends heavily on what functions we are optimizing, and the convergence gap between optimal and random permutations can vary from exponential to nonexistent. We first show that for 1-dimensional strongly convex functions, with smooth second derivatives, there exist optimal permutations that offer exponentially faster convergence compared to random. However, for general strongly convex functions, random permutations are optimal. Finally, we show that for quadratic, strongly-convex functions, there are easy-to-construct permutations that lead to accelerated convergence compared to random. Our results suggest that a general convergence characterization of optimal permutations cannot capture the nuances of individual function classes, and can mistakenly indicate that one cannot do much better than random.

****

Comparing Distributions by Measuring Differences that Affect Decision Making
Shengjia Zhao,Abhishek Sinha,Yutong He,Aidan Perreault,Jiaming Song,Stefano Ermon

Measuring the discrepancy between two probability distributions is a fundamental problem in machine learning and statistics. We propose a new class of discrepancies based on the optimal loss for a decision task -- two distributions are different if the optimal decision loss is higher on their mixture than on each individual distribution. By suitably choosing the decision task, this generalizes the Jensen-Shannon divergence and the maximum mean discrepancy family. We apply our approach to two-sample tests, and on various benchmarks, we achieve superior test power compared to competing methods. In addition, a modeler can directly specify their preferences when comparing distributions through the decision loss. We apply this property to understanding the effects of climate change on different social and economic activities, evaluating sample quality, and selecting features targeting different decision tasks.

****

Message Passing Neural PDE Solvers

Johannes Brandstetter,Daniel E. Worrall,Max Welling

The numerical solution of partial differential equations (PDEs) is difficult, having led to a century of research so far. Recently, there have been pushes to build neural--numerical hybrid solvers, which piggy-backs the modern trend towards fully end-to-end learned systems. Most works so far can only generalize over a subset of properties to which a generic solver would be faced, including: resolution, topology, geometry, boundary conditions, domain discretization regularity, dimensionality, etc. In this work, we build a solver, satisfying these properties, where all the components are based on neural message passing, replacing all heuristically designed components in the computation graph with backprop-optimized neural function approximators. We show that neural message passing solvers representationally contain some classical methods, such as finite differences, finite volumes, and WENO schemes. In order to encourage stability in training autoregressive models, we put forward a method that is based on the principle of zero-stability, posing stability as a domain adaptation problem. We validate our method on various fluid-like flow problems, demonstrating fast, stable, and accurate performance across different domain topologies, discretization, etc. in 1D and 2D. Our model outperforms state-of-the-art numerical solvers in the low resolution regime in terms of speed, and accuracy.

****

Graph-less Neural Networks: Teaching Old MLPs New Tricks Via Distillation

Shichang Zhang,Yozen Liu,Yizhou Sun,Neil Shah

Graph Neural Networks (GNNs) are popular for graph machine learning and have shown great results on wide node classification tasks. Yet, they are less popular for practical deployments in the industry owing to their scalability challenges incurred by data dependency. Namely, GNN inference depends on neighbor nodes multiple hops away from the target, and fetching them burdens latency-constrained applications. Existing inference acceleration methods like pruning and quantization can speed up GNNs by reducing Multiplication-and-ACcumulation (MAC) operations, but the improvements are limited given the data dependency is not resolved. Conversely, multi-layer perceptrons (MLPs) have no graph dependency and infer much faster than GNNs, even though they are less accurate than GNNs for node classification in general. Motivated by these complementary strengths and weaknesses, we bring GNNs and MLPs together via knowledge distillation (KD). Our work shows that the performance of MLPs can be improved by large margins with GNN KD. We call the distilled MLPs Graph-less Neural Networks (GLNNs) as they have no inference graph dependency. We show that GLNNs with competitive accuracy infer faster than GNNs by 146X-273X and faster than other acceleration methods by 14X-27X. Under a production setting involving both transductive and inductive predictions across 7 datasets, GLNN accuracies improve over stand-alone MLPs by 12.36% on average and match GNNs on 6/7 datasets. Comprehensive analysis shows when and why GLNNs can achieve competitive accuracies to GNNs and suggests GLNN as a handy choice for latency-constrained applications.

****

Provably Calibrated Regression Under Distribution Drift

Shengjia Zhao,YUSUKE TASHIRO,Danny Tse,Stefano Ermon

Accurate uncertainty quantification is a key building block of trustworthy machine learning systems. Uncertainty is typically represented by probability distributions over the possible outcomes, and these probabilities should be calibrated, \textit{e.g}. the 90\% credible interval should contain the true outcome 90\% of the times. In the online prediction setup, existing conformal methods can provably achieve calibration assuming no distribution shift; however, the assumption is difficult to verify, and unlikely to hold in many applications such as time series prediction. Inspired by control theory, we propose a prediction algorithm that guarantees calibration even under distribution shift, and achieves strong performance on metrics such as sharpness and proper scores. We compare our method with baselines on 19 time-series and regression datasets, and our method achieves approximately 2x reduction in calibration error, comparable sharpness, and i mproved downstream decision utility.

****

Density Estimation for Conservative Q-Learning
Paul Daoudi,Merwan Barlier,Ludovic Dos Santos,Aladin Virmaux
Batch Reinforcement Learning algorithms aim at learning the best policy from a batch of data without interacting with the environment. Within this setting, one difficulty is to correctly assess the value of state-action pairs that are far from the dataset. Indeed, the lack of information may provoke an overestimation of the value function, leading to non-desirable behaviors. A compromise between enhancing the behaviour policy's performance and staying close to it must be found. To alleviate this issue, most existing approaches introduce a regularization term to favor state-action pairs from the dataset. In this paper, we refine this idea by estimating the density of these state-action pairs to distinguish neighbourhoods. The resulting regularization guides the policy toward meaningful unseen regions, improving the learning process. We hence introduce Density Conservative Q-Learning (D-CQL), a batch-RL algorithm with strong theoretical guarantees that carefully penalizes the value function based on the amount of information collected in the state-action space. The performance of our approach is outlined on many classical benchmark in batch-RL.

****

Cost-Sensitive Hierarchical Classification through Layer-wise Abstentions
Alycia Lee,Anthony L Pineci,Uriah Israel,Omer Bar-Tal,Leeat Keren,David A. Van Valen,Anima Anandkumar,Yisong Yue,Anqi Liu
We study the problem of cost-sensitive hierarchical classification where a label taxonomy has a cost-sensitive loss associated with it, which represents the cost of (wrong) predictions at different levels of the hierarchy. Directly optimizing the cost-sensitive hierarchical loss is hard, due to its non-convexity, especially when the size of the taxonomy is large. In this paper, we propose a \textbf{L}ayer-wise \textbf{A}bstaining Loss \textbf{M}inimization method (LAM), a tractable method that breaks the hierarchical learning problem into layer-by-layer learning-to-abstain sub-problems. We prove that there is a bijective mapping between the original hierarchical cost-sensitive loss and the set of layer-wise abstaining losses under symmetry assumptions. We employ the distributionally robust learning framework to solve the learning-to-abstain problems in each layer. We conduct experiments on large-scale bird dataset as well as on cell classification problems. Our results demonstrate that LAM achieves a lower hierarchical cost-sensitive loss in high accuracy regions, compared to previous methods and their modified versions for a fair comparison, even though they are not directly optimizing this loss. For each layer, we also achieve higher accuracy when the overall accuracy is kept fixed across different methods. Furthermore, we also show the flexibility of LAM by proposing a per-class loss-adjustment heuristic to achieve a performance profile. This can be used for cost design to translate user requirements into optimizable cost functions.

****

A Koopman Approach to Understanding Sequence Neural Models
Ilan Naiman,Omri Azencot
Deep learning models are often treated as "black boxes". Existing approaches for understanding the decision mechanisms of neural networks provide limited explanations or depend on local theories. Recently, a data-driven framework based on Koopman theory was developed for the analysis of nonlinear dynamical systems. In this paper, we introduce a new approach to understanding trained sequence neural models: the Koopman Analysis of Neural Networks (KANN) method. At the core of our method lies the Koopman operator, which is linear, yet it encodes the dominant features of the network latent dynamics. Moreover, its eigenvectors and eigenvalues facilitate understanding: in the sentiment analysis problem, the eigenvectors highlight positive and negative n-grams; and, in the ECG classification challenge, the eigenvectors capture the dominant features of the normal beat signal.

****

Temporal abstractions-augmented temporally contrastive learning: an alternative to the Laplacian in RL

Akram Erraqabi,Marlos C. Machado,Mingde Zhao,Sainbayar Sukhbaatar,Ludovic Denoyer,Alessandro Lazaric,Yoshua Bengio

In reinforcement learning (RL), the graph Laplacian has proved to be a valuable tool in the task-agnostic setting, with applications ranging from option discovery to dynamics-aware metric learning. Conveniently, learning the Laplacian representation has recently been framed as the optimization of a temporally-contrastive objective to overcome its computational limitations in large or even continuous state spaces (Wu et al., 2019). However, this approach relies on a uniform access to the state space S, and overlooks the exploration problem that emerges during the representation learning process. In this work, we reconcile such representation learning with exploration in a non-uniform prior setting, while recovering the expressive potential afforded by a uniform prior. Our approach leverages the learned representation to build a skill-based covering policy which in turn provides a better training distribution to extend and refine the representation. We also propose to integrate temporal abstractions captured by the learned skills into the representation, which encourages exploration and improves the representation’s dynamics-awareness. We find that our method scales better to challenging environments, and that the learned skills can solve difficult continuous navigation tasks with sparse rewards, where standard skill discovery methods are limited.

****

Deep Classifiers with Label Noise Modeling and Distance Awareness

Vincent Fortuin,Mark Collier,Florian Wenzel,James Urquhart Allingham,Jeremiah Zhe Liu,Dustin Tran,Balaji Lakshminarayanan,Jesse Berent,Rodolphe Jenatton,Effrosyni Kokiopoulou

Uncertainty estimation in deep learning has recently emerged as a crucial area of interest to advance reliability and robustness in safety-critical applications. While there have been many proposed methods that either focus on distance-aware model uncertainties for out-of-distribution detection or on input-dependent label uncertainties for in-distribution calibration, both of these types of uncertainty are often necessary. In this work, we propose the HetSNGP method for jointly modeling the model and data uncertainty. We show that our proposed model affords a favorable combination between these two complementary types of uncertainty and thus outperforms the baseline methods on some challenging out-of-distribution datasets, including CIFAR-100C, Imagenet-C, and Imagenet-A. Moreover, we propose HetSNGP Ensemble, an ensembled version of our method which adds an additional type of uncertainty and also outperforms other ensemble baselines.

****

Relating transformers to models and neural representations of the hippocampal formation

James C. R. Whittington,Joseph Warren,Tim E.J. Behrens

Many deep neural network architectures loosely based on brain networks have recently been shown to replicate neural firing patterns observed in the brain. One of the most exciting and promising novel architectures, the Transformer neural network, was developed without the brain in mind. In this work, we show that transformers, when equipped with recurrent position encodings, replicate the precisely tuned spatial representations of the hippocampal formation; most notably place and grid cells. Furthermore, we show that this result is no surprise since it is closely related to current hippocampal models from neuroscience. We additionally show the transformer version offers dramatic performance gains over the neuroscience version. This work continues to bind computations of artificial and brain networks, offers a novel understanding of the hippocampal-cortical interaction, and suggests how wider cortical areas may perform complex tasks beyond current neuroscience models such as language comprehension.

****

Sequential Covariate Shift Detection Using Classifier Two-Sample Tests

Sooyong Jang,Sangdon Park,Insup Lee,Osbert Bastani

A standard assumption in supervised learning is that the training data and test data are from the same distribution. However, this assumption often fails to hold in practice, which can cause the learned model to perform poorly. We consider the problem of detecting covariate shift, where the covariate distribution shifts but the conditional distribution of labels given covariates remains the same. This problem can naturally be solved using a two-sample test--- i.e., test whether the current test distribution of covariates equals the training distribution of covariates. Our algorithm builds on classifier tests, which train a discriminator to distinguish train and test covariates, and then use the accuracy of this discriminator as a test statistic. A key challenge is that classifier tests assume given a fixed set of test covariates. In practice, test covariates often arrive sequentially over time---e.g., a self-driving car observes a stream of images while driving. Furthermore, covariate shift can occur multiple times--- i.e., shift and then shift back later or gradually shift over time. To address these challenges, our algorithm trains the discriminator online. Furthermore, it evaluates test accuracy using each new covariate before taking a gradient step; this strategy avoids constructing a held-out test set, which can reduce sample efficiency. We prove that this optimization preserves the correctness---i.e., our algorithm achieves a desired bound on the false positive rate. In our experiments, we show that our algorithm efficiently detects covariate shifts on ImageNet.

****

Provable hierarchical lifelong learning with a sketch-based modular architecture

Rina Panigrahy,Brendan Juba,Zihao Deng,Xin Wang,Zee Fryer

We propose a modular architecture for lifelong learning of hierarchically structured tasks. Specifically, we prove that our architecture is theoretically able to learn tasks that can be solved by functions that are learnable given access to functions for other, previously learned tasks as subroutines. We show that some tasks that we can learn in this way are not learned by standard training methods in practice; indeed, prior work suggests that some such tasks cannot be learned by \emph{any} efficient method without the aid of the simpler tasks. We also consider methods for identifying the tasks automatically, without relying on explicitly given indicators.

****

How many degrees of freedom do we need to train deep networks: a loss landscape perspective

Brett W Larsen,Stanislav Fort,Nic Becker,Surya Ganguli

A variety of recent works, spanning pruning, lottery tickets, and training within random subspaces, have shown that deep neural networks can be trained using far fewer degrees of freedom than the total number of parameters. We analyze this phenomenon for random subspaces by first examining the success probability of hitting a training loss sublevel set when training within a random subspace of a given training dimensionality. We find a sharp phase transition in the success probability from 0 to 1 as the training dimension surpasses a threshold. This threshold training dimension increases as the desired final loss decreases, but decreases as the initial loss decreases. We then theoretically explain the origin of this phase transition, and its dependence on initialization and final desired loss, in terms of properties of the high-dimensional geometry of the loss landscape. In particular, we show via Gordon's escape theorem, that the training dimension plus the Gaussian width of the desired loss sublevel set, projected onto a unit sphere surrounding the initialization, must exceed the total number of parameters for the success probability to be large. In several architectures and datasets, we measure the threshold training dimension as a function of initialization and demonstrate that it is a small fraction of the total parameters, implying by our theory that successful training with so few dimensions is possible precisely because the Gaussian width of low loss sublevel sets is very large. Moreover, we compare this threshold training dimension to more sophisticated ways of reducing training degrees of freedom, including lottery tickets as well as a new, analogous method: lottery subspaces.

****

Is Importance Weighting Incompatible with Interpolating Classifiers?

Ke Alexander Wang,Niladri Shekhar Chatterji,Saminul Haque,Tatsunori Hashimoto

Importance weighting is a classic technique to handle distribution shifts. However, prior work has presented strong empirical and theoretical evidence demonstra ting that importance weights can have little to no effect on overparameterized neural networks. \emph{Is importance weighting truly incompatible with the training of overparameterized neural networks?} Our paper answers this in the negative. We show that importance weighting fails not because of the overparameterization, but instead, as a result of using exponentially-tailed losses like the logistic or cross-entropy loss. As a remedy, we show that polynomially-tailed losses restore the effects of importance reweighting in correcting distribution shift in overparameterized models. We characterize the behavior of gradient descent on importance weighted polynomially-tailed losses with overparameterized linear models, and theoretically demonstrate the advantage of using polynomially-tailed losses in a label shift setting. Surprisingly, our theory shows that using weights that are obtained by exponentiating the classical unbiased importance weights can improve performance. Finally, we demonstrate the practical value of our analysis with neural network experiments on a subpopulation shift and a label shift dataset. When reweighted, our loss function can outperform reweighted cross-entropy by as much as 9\% in test accuracy. Our loss function also gives test accuracies comparable to, or even exceeding, well-tuned state-of-the-art methods for correcting distribution shifts.

****

MAGNEx: A Model Agnostic Global Neural Explainer

Nikolaos Manginas,Prodromos Malakasiotis,Eirini Spyropoulou,Ion Androutsopoulos,Georgios Paliouras

Black-box decision models have been widely adopted both in industry and academia due to their excellent performance across many challenging tasks and domains. However, much criticism has been raised around modern AI systems, to a large extent due to their inability to produce explainable decisions that both their end-users and their developers can trust. The need for such decisions, i.e., decisions accompanied by a rationale for why they are made, has ignited much recent research. We propose MAGNEx, a global algorithm that leverages neural-network based explainers to produce rationales for any black-box decision model, neural or not. MAGNEx is model-agnostic, and thus easily generalizable across domains and applications. More importantly, MAGNEx is global, i.e., it learns to create rationales by optimizing for a number of instances at once, contrary to local methods that aim at explaining a single example. The global nature of MAGNEx has two advantages over local methods: i) it generalizes across instances hence producing more faithful explanations, ii) it is computationally more efficient during inference. Our experiments confirm that MAGNEx outperforms popular explainability algorithms both in explanation quality and in computational efficiency.

****

A Simple Reward-free Approach to Constrained Reinforcement Learning

Sobhan Miryoosefi,Chi Jin

In constrained reinforcement learning (RL), a learning agent seeks to not only optimize the overall reward but also satisfy the additional safety, diversity, or budget constraints. Consequently, existing constrained RL solutions require several new algorithmic ingredients that are notably different from standard RL. On the other hand, reward-free RL is independently developed in the unconstrained literature, which learns the transition dynamics without using the reward information, and thus naturally capable of addressing RL with multiple objectives under the common dynamics. This paper bridges reward-free RL and constrained RL. Particularly, we propose a simple meta-algorithm such that given any reward-free RL oracle, the approachability and constrained RL problems can be directly solved with negligible overheads in sample complexity. Utilizing the existing reward-free RL solvers, our framework provides sharp sample complexity results for constrained RL in the tabular MDP setting, matching the best existing results up to a factor of horizon dependence; our framework directly extends to a setting of tabular two-player Markov games, and gives a new result for constrained RL with linear function approximation.

****

Learning and controlling the source-filter representation of speech with a variational autoencoder

Samir Alain Sadok,Simon Leglaive,Laurent Girin,Xavier Alameda-Pineda,Renaud Séguier

Understanding and controlling latent representations in deep generative models is a challenging yet important problem for analyzing, transforming and generating various types of data. In speech processing, inspiring from the anatomical mechanisms of phonation, the source-filter model considers that speech signals are produced from a few independent and physically meaningful continuous latent factors, among which the fundamental frequency and the formants are of primary importance. In this work, we show that the source-filter model of speech production naturally arises in the latent space of a variational autoencoder (VAE) trained in an unsupervised fashion on a dataset of natural speech signals. Using speech signals generated with an artificial speech synthesizer, we experimentally demonstrate that the fundamental frequency and formant frequencies are encoded in orthogonal subspaces of the VAE latent space and we develop a weakly-supervised method to accurately and independently control these speech factors of variation within the learned latent subspaces. Without requiring additional information such as text or human-labeled data, we propose a deep generative model of speech spectrograms that is conditioned on the fundamental frequency and formant frequencies, and which is applied to the transformation of speech signals.

****

Neural Models for Output-Space Invariance in Combinatorial Problems

Yatin Nandwani,Vidit Jain,Mausam .,Parag Singla

Recently many neural models have been proposed to solve combinatorial puzzles by implicitly learning underlying constraints using their solved instances, such as sudoku or graph coloring (GCP). One drawback of the proposed architectures, which are often based on Graph Neural Networks (GNN) (Zhou et al., 2020), is that they cannot generalize across the size of the output space from which variables are assigned a value, for example, set of colors in a GCP, or board-size in sudoku. We call the output space for the variables as ‘value-set’. While many works have demonstrated generalization of GNNs across graph size, there has been no study on how to design a GNN for achieving value-set invariance for problems that come from the same domain. For example, learning to solve 16 x 16 sudoku after being trained on only 9 x 9 sudokus, or coloring a 7 colorable graph after training on 4 colorable graphs. In this work, we propose novel methods to extend GNN based architectures to achieve value-set invariance. Specifically, our model builds on recently proposed Recurrent Relational Networks (RRN) (Palm et al., 2018). Our first approach exploits the graph-size invariance of GNNs by converting a multi-class node classification problem into a binary node classification problem. Our second approach works directly with multiple classes by adding multiple nodes corresponding to the values in the value-set, and then connecting variable nodes to value nodes depending on the problem initialization. Our experimental evaluation on three different combinatorial problems demonstrates that both our models perform well on our novel problem, compared to a generic neural reasoner. Between two of our models, we observe an inherent trade-off: while the binarized model gives better performance when trained on smaller value-sets, multi-valued model is much more memory efficient, resulting in improved performance when trained on larger value-sets, where binarized model fails to train.

****

StyleNeRF: A Style-based 3D Aware Generator for High-resolution Image Synthesis

Jiatao Gu,Lingjie Liu,Peng Wang,Christian Theobalt

We propose StyleNeRF, a 3D-aware generative model for photo-realistic high-resolution image synthesis with high multi-view consistency, which can be trained on unstructured 2D images. Existing approaches either cannot synthesize high-resolution images with fine details or yield clearly noticeable 3D-inconsistent artifacts. In addition, many of them lack control on style attributes and explicit 3D camera poses. To address these issues, StyleNeRF integrates the neural radiance field (NeRF) into a style-based generator to tackle the aforementioned challenges, i.e., improving rendering efficiency and 3D consistency for high-resolution image generation. To address the first issue, we perform volume rendering only to produce a low-resolution feature map, and progressively apply upsampling in 2D

. To mitigate the inconsistencies caused by 2D upsampling, we propose multiple designs including a better upsampler choice and a new regularization loss to enforce 3D consistency. With these designs, StyleNeRF is able to synthesize high-resolution images at interactive rates while preserving 3D consistency at high quality. StyleNeRF also enables control of camera poses and different levels of styles, which can generalize to unseen views. It also supports challenging tasks such as style mixing, inversion and simple semantic edits.

****

Learning Context-Adapted Video-Text Retrieval by Attending to User Comments
Laura Hanu,Yuki M Asano,James Thewlis,Christian Rupprecht
Learning strong representations for multi-modal retrieval is an important problem for many applications, such as recommendation and search. Current benchmarks and even datasets are often manually constructed and consist of mostly clean samples where all modalities are well-correlated with the content. Thus, current video-text retrieval literature largely focuses on video titles or audio transcripts, while ignoring user comments, since users often tend to discuss topics only vaguely related to the video.
In this paper we present a novel method that learns meaningful representations from videos, titles and comments, which are abundant on the internet. Due to the nature of user comments, we introduce an attention-based mechanism that allows the model to disregard text with irrelevant content.
In our experiments, we demonstrate that, by using comments, our method is able to learn better, more contextualised, representations, while also achieving competitive results on standard video-text retrieval benchmarks.

****

Resilience to Multiple Attacks via Adversarially Trained MIMO Ensembles
Ruqi Bai,David I. Inouye,Saurabh Bagchi
While ensemble methods have been widely used for robustness against random perturbations (\ie the average case), ensemble approaches for robustness against adversarial perturbations (\ie the worst case) have remained elusive despite multiple prior attempts. We show that ensemble methods can improve adversarial robustness to multiple attacks if the ensemble is \emph{adversarially diverse}, which is defined by two properties: 1) the sub-models are adversarially robust themselves and yet 2) adversarial attacks do not transfer easily between sub-models. While at first glance, creating such an ensemble would seem computationally expensive, we demonstrate that an adversarially diverse ensemble can be trained with minimal computational overhead via a Multiple-Input Multiple-Output (MIMO) model. Specifically, we propose to train a MIMO model with adversarial training ({\emph{MAT}}), where each sub-model can be trained on a different attack type. When computing gradients for generating adversarial examples during training, we use the gradient with respect to the ensemble objective. This has a two-fold benefit: 1) it only requires 1 backward pass and 2) the cross-gradient information between the models promotes robustness against transferable attacks. We empirically demonstrate that {\emph{MAT}} produces an ensemble of models that is adversarially diverse and significantly improves performance over single models or vanilla ensembles while being comparable to previous state-of-the-art methods. On MNIST, we obtain 99.5% clean accuracy and ($88.6\%, 57.1\%,71.6\%$) against $(\ell_\infty, \ell_2, \ell_1)$ attacks, and on CIFAR10, we achieve 79.7% clean accuracy and ($47.9\%, 61.8\%,47.6\%$) against $(\ell_\infty, \ell_2, \ell_1)$ attacks, which are comparable to previous state-of-the-art methods.

****

Offline-Online Reinforcement Learning: Extending Batch and Online RL
Maryam Hashemzadeh,Wesley Chung,Martha White
Batch RL has seen a surge in popularity and is applicable in many practical scenarios where past data is available. Unfortunately, the performance of batch RL agents is limited in both theory and practice without strong assumptions on the data-collection process e.g. sufficient coverage or a good policy. To enable better performance, we investigate the offline-online setting: The agent has access

to a batch of data to train on but is also allowed to learn during the evaluation phase in an online manner. This is an extension to batch RL, allowing the agent to adapt to new situations without having to precommit to a policy. In our experiments, we find that agents trained in an offline-online manner can outperform agents trained only offline or online, sometimes by a large margin, for different dataset sizes and data-collection policies. Furthermore, we investigate the use of optimism vs. pessimism for value functions in the offline-online setting due to their use in batch and online RL.

****

The Role of Pretrained Representations for the OOD Generalization of RL Agents

Frederik Träuble,Andrea Dittadi,Manuel Wuthrich,Felix Widmaier,Peter Vincent Gehler,Ole Winther,Francesco Locatello,Olivier Bachem,Bernhard Schölkopf,Stefan Bauer

Building sample-efficient agents that generalize out-of-distribution (OOD) in real-world settings remains a fundamental unsolved problem on the path towards achieving higher-level cognition. One particularly promising approach is to begin with low-dimensional, pretrained representations of our world, which should facilitate efficient downstream learning and generalization. By training 240 representations and over 10,000 reinforcement learning (RL) policies on a simulated robotic setup, we evaluate to what extent different properties of pretrained VAE-based representations affect the OOD generalization of downstream agents. We observe that many agents are surprisingly robust to realistic distribution shifts, including the challenging sim-to-real case. In addition, we find that the generalization performance of a simple downstream proxy task reliably predicts the generalization performance of our RL agents under a wide range of OOD settings. Such proxy tasks can thus be used to select pretrained representations that will lead to agents that generalize.

****

On the Latent Holes ■ of VAEs for Text Generation

Ruizhe Li,Xutan Peng,Chenghua Lin

In this paper, we provide the first focused study on the discontinuities (aka. holes) in the latent space of Variational Auto-Encoders (VAEs), a phenomenon which has been shown to have a detrimental effect on model capacity. When investigating la- tent holes, existing works are exclusively centred around the encoder network and they merely explore the existence of holes. We tackle these limitations by proposing a highly efficient Tree-based Decoder-Centric (TDC) algorithm for latent hole identification, with a focal point on the text domain. In contrast to past studies, our approach pays attention to the decoder network, as a decoder has a direct impact on the model’s output quality. Furthermore, we provide, for the first time, in-depth empirical analysis of the latent hole phenomenon, investigating several important aspects such as how the holes impact VAE algorithms’ performance on text generation, and how the holes are distributed in the latent space.

****

Learning Minimal Representations with Model Invariance

Manan Tomar,Amy Zhang,Matthew E. Taylor

Sparsity has been identified as an important characteristic in learning neural networks that generalize well, forming the key idea in constructing minimal representations. Minimal representations are ones that only encode information required to predict well on a task and nothing more. In this paper we present a powerful approach to learning minimal representations. Our method, called ModInv or model invariance, argues for learning using multiple predictors and a single representation, creating a bottleneck architecture. Predictors' learning landscapes are diversified by training independently and with different learning rates. The common representation acts as a implicit invariance objective to avoid the different spurious correlations captured by individual predictors. This in turn leads to better generalization performance. ModInv is tested on both the Reinforcement Learning and the Self-supervised Learning settings, showcasing strong performance boosts in both. It is extremely simple to implement, does not lead to any delay in walk clock times while training, and can be applied across different pro blem settings.

****

Show Your Work: Scratchpads for Intermediate Computation with Language Models

Maxwell Nye,Anders Johan Andreassen,Guy Gur-Ari,Henryk Michalewski,Jacob Austin,David Bieber,David Dohan,Aitor Lewkowycz,Maarten Bosma,David Luan,Charles Sutton,Augustus Odena

Large pre-trained language models perform remarkably well on tasks that can be done "in one pass", such as generating realistic text or synthesizing computer programs. However, they struggle with tasks that require unbounded multi-step computation, such as adding integers or executing programs. Surprisingly, we find that these same models are able to perform complex multi-step computations --- even in the few-shot regime --- when asked to perform the operation "step by step", showing the results of intermediate computations. In particular, we train transformers to perform multi-step computations by asking them to emit intermediate computation steps into a "scratchpad". On a series of increasingly complex tasks ranging from long addition to the execution of arbitrary programs, we show that scratchpads dramatically improve the ability of language models to perform multi-step computations.

****

Back to Basics: Efficient Network Compression via IMP

Max Zimmer,Sebastian Pokutta,Christoph Spiegel

Network pruning is a widely used technique for effectively compressing Deep Neural Networks with little to no degradation in performance during inference. Iterative Magnitude Pruning (IMP) (Han et al., 2015) is one of the most established approaches for network pruning, consisting of several iterative training and pruning steps, where a significant amount of the network’s performance is lost after pruning and then recovered in the subsequent retraining phase. While commonly used as a benchmark reference, it is often argued that a) it reaches suboptimal states by not incorporating sparsification into the training phase, b) its global selection criterion fails to properly determine optimal layer-wise pruning rates and c) its iterative nature makes it slow and non-competitive. In light of recently proposed retraining techniques, we investigate these claims through rigorous and consistent experiments where we compare IMP to pruning-during-training algorithms, evaluate proposed modifications of its selection criterion and study the number of iterations and total training time actually required. We find that IMP with SLR (Le & Hua, 2021) for retraining can outperform state-of-the-art pruning-during-training approaches without or with only little computational overhead, that the global magnitude selection criterion is largely competitive with more complex approaches and that only few retraining epochs are needed in practice to achieve most of the sparsity-vs.-performance trade-off of IMP. Our goals are both to demonstrate that basic IMP can already provide state-of-the-art pruning results on par or outperforming more complex or heavily parameterized approaches and also to establish a more realistic yet easily realisable baseline for future research.

****

Learning Representations for Pixel-based Control: What Matters and Why?

Manan Tomar,Utkarsh Aashu Mishra,Amy Zhang,Matthew E. Taylor

Learning representations for pixel-based control has garnered significant attention recently in reinforcement learning. A wide range of methods have been proposed to enable efficient learning, leading to sample complexities similar to those in the full state setting. However, moving beyond carefully curated pixel data sets (centered crop, appropriate lighting, clear background, etc.) remains challenging. In this paper, we adopt a more difficult setting, incorporating background distractors, as a first step towards addressing this challenge. We present a simple baseline approach that can learn meaningful representations with no metric-based learning, no data augmentations, no world-model learning, and no contrastive learning. We then analyze when and why previously proposed methods are likely to fail or reduce to the same performance as the baseline in this harder setting and why we should think carefully about extending such methods beyond the well-curated environments. Our results show that finer categorization of benchmark s on the basis of characteristics like the density of reward, planning horizon of the problem, presence of task-irrelevant components, etc., is crucial in evaluating algorithms. Based on these observations, we propose different metrics to consider when evaluating an algorithm on benchmark tasks. We hope such a data-centric view can motivate researchers to rethink representation learning when investigating how to best apply RL to real-world tasks.

****

SGDEM: stochastic gradient descent with energy and momentum

Hailiang Liu,Xuping Tian

In this paper, we propose SGDEM, Stochastic Gradient Descent with Energy and Momentum to solve a large class of general nonconvex stochastic optimization problems, based on the AEGD method that originated in the work [AEGD: Adaptive Gradient Descent with Energy. arXiv: 2010.05109]. SGDEM incorporates both energy and momentum at the same time so as to inherit their dual advantages. We show that SGDEM features an unconditional energy stability property, and derive energy-dependent convergence rates in the general nonconvex stochastic setting, as well as a regret bound in the online convex setting. A lower threshold for the energy variable is also provided. Our experimental results show that SGDEM converges faster than AEGD and generalizes better or at least as well as SGDM in training some deep neural networks.

****

Learning to Act with Affordance-Aware Multimodal Neural SLAM

Zhiwei Jia,Kaixiang Lin,Yizhou Zhao,Qiaozi Gao,Govind Thattai,Gaurav S. Sukhatme

Recent years have witnessed an emerging paradigm shift toward embodied artificial intelligence, in which an agent must learn to solve challenging tasks by interacting with its environment. There are several challenges in solving embodied multimodal tasks, including long-horizon planning, vision-and-language grounding, and efficient exploration. We focus on a critical bottleneck, namely the performance of planning and navigation. To tackle this challenge, we propose a Neural SLAM approach that, for the first time, utilizes several modalities for exploration, predicts an affordance-aware semantic map, and plans over it at the same time. This significantly improves exploration efficiency, leads to robust long-horizon planning, and enables effective vision-and-language grounding. With the proposed Affordance-aware Multimodal Neural SLAM (AMSLAM) approach, we obtain more than 40% improvement over prior published work on the ALFRED benchmark and set a new state-of-the-art generalization performance at a success rate of 23.48% on the test unseen scenes.

****

Enabling Arbitrary Translation Objectives with Adaptive Tree Search

Wang Ling,Wojciech Stokowiec,Domenic Donato,Chris Dyer,Lei Yu,Laurent Sartran,Austin Matthews

We introduce an adaptive tree search algorithm, which is a deterministic variant of Monte Carlo tree search, that can find high-scoring outputs under translation models that make no assumptions about the form or structure of the search objective. This algorithm enables the exploration of new kinds of models that are unencumbered by constraints imposed to make decoding tractable, such as autoregressivity or conditional independence assumptions. When applied to autoregressive models, our algorithm has different biases than beam search has, which enables a new analysis of the role of decoding bias in autoregressive models. Empirically, we show that our adaptive tree search algorithm finds outputs with substantially better model scores compared to beam search in autoregressive models, and compared to reranking techniques in models whose scores do not decompose additively with respect to the words in the output. We also characterise the correlation of several translation model objectives with respect to BLEU. We find that while some standard models are poorly calibrated and benefit from the beam search bias, other often more robust models (autoregressive models tuned to maximize expected automatic metric scores, the noisy channel model and a newly proposed objective) benefit from increasing amounts of search using our proposed decoder, whereas the beam search bias limits the improvements obtained from such objectives. Thus, we argue that as models improve, the improvements may be masked by over-relia nce on beam search or reranking based methods.

****

Neural Capacitance: A New Perspective of Neural Network Selection via Edge Dynamics
Chunheng Jiang,Tejaswini Pedapati,Pin-Yu Chen,Yizhou Sun,Jianxi Gao
Efficient model selection for identifying a suitable pre-trained neural network to a downstream task is a fundamental yet challenging task in deep learning. Current practice requires expensive computational costs in model training for performance prediction. In this paper, we propose a novel framework for neural network selection by analyzing the governing dynamics over synaptic connections (edges) during training. Our framework is built on the fact that back-propagation during neural network training is equivalent to the dynamical evolution of synaptic connections. Therefore, a converged neural network is associated with an equilibrium state of a networked system composed of those edges. To this end, we construct a network mapping ϕ, converting a neural network G_A to a directed line graph G_B that is defined on those edges in G_A. Next, we derive a \textit{neural capacitance} metric $\beta_{\rm eff}$ as a predictive measure universally capturing the generalization capability of G_A on the downstream task using only a handful of early training results. We carried out extensive experiments using 17 popular pre-trained ImageNet models and five benchmark datasets, including CIFAR10, CIFAR100, SVHN, Fashion MNIST and Birds, to evaluate the fine-tuning performance of our framework. Our neural capacitance metric is shown to be a powerful indicator for model selection based only on early training results and is more efficient than state-of-the-art methods.

****

The Evolution of Out-of-Distribution Robustness Throughout Fine-Tuning
Anders Johan Andreassen,Yasaman Bahri,Behnam Neyshabur,Rebecca Roelofs
Although machine learning models typically experience a drop in performance on out-of-distribution data, accuracies on in- versus out-of-distribution data are widely observed to follow a single linear trend when evaluated across a testbed of models. Models that are more accurate on the out-of-distribution data relative to this baseline exhibit “effective robustness” and are exceedingly rare. Identifying such models, and understanding their properties, is key to improving out-of-distribution performance. We conduct a thorough empirical investigation of effective robustness during fine-tuning and surprisingly find that models pre-trained on larger datasets exhibit effective robustness during training that vanishes at convergence. We study how properties of the data influence effective robustness, and we show that it increases with the larger size, more diversity, and higher example difficulty of the dataset. We also find that models that display effective robustness are able to correctly classify 10% of the examples that no other current testbed model gets correct. Finally, we discuss several strategies for scaling effective robustness to the high-accuracy regime to improve the out-of-distribution accuracy of state-of-the-art models.

****

f-Divergence Thermodynamic Variational Objective: a Deformed Geometry Perspective
Jun Li,Ping Li
In this paper, we propose a f-divergence Thermodynamic Variational Objective (f-TVO). f-TVO generalizes the Thermodynamic Variational Objective (TVO) by replacing Kullback–Leibler (KL) divergence with arbitary differeitiable f-divergence. In particular, f-TVO approximates dual function of model evidence $f^*(p(x))$ rather than the log model evidence $\log p(x)$ in TVO. f-TVO is derived from a deformed χ-geometry perspective. By defining χ-exponential family exponential, we are able to integral f-TVO along the χ-path, which is the deformed geodesic between variational posterior distribution and true posterior distribution. Optimizing scheme of f-TVO includes reparameterization trick and Monte Carlo approximation. Experiments on VAE and Bayesian neural network show that the proposed f-TVO performs better than cooresponding baseline f-divergence variational inference.

****

Multi-Stage Episodic Control for Strategic Exploration in Text Games

Jens Tuyls,Shunyu Yao,Sham M. Kakade,Karthik R Narasimhan

Text adventure games present unique challenges to reinforcement learning methods due to their combinatorially large action spaces and sparse rewards. The interplay of these two factors is particularly demanding because large action spaces require extensive exploration, while sparse rewards provide limited feedback. This work proposes to tackle the explore-vs-exploit dilemma using a multi-stage approach that explicitly disentangles these two strategies within each episode. Our algorithm, called eXploit-Then-eXplore (XTX), begins each episode using an exploitation policy that imitates a set of promising trajectories from the past, and then switches over to an exploration policy aimed at discovering novel actions that lead to unseen state spaces. This policy decomposition allows us to combine global decisions about which parts of the game space to return to with curiosity-based local exploration in that space, motivated by how a human may approach these games. Our method significantly outperforms prior approaches by 27% and 11% average normalized score over 12 games from the Jericho benchmark (Hausknecht et al., 2020) in both deterministic and stochastic settings, respectively. On the game of Zork1, in particular, XTX obtains a score of 103, more than a 2x improvement over prior methods, and pushes past several known bottlenecks in the game that have plagued previous state-of-the-art methods.

****

Squeezing SGD Parallelization Performance in Distributed Training Using Delayed Averaging

Pengcheng Li,Yixin Guo,Yawen Zhang,Qinggang Zhou

State-of-the-art deep learning algorithms rely on distributed training to tackle the increasing model size and training data. Mini-batch Stochastic Gradient Descent (SGD) requires workers to halt forward/backward propagations, to wait for gradients synchronized among all workers before the next batch of tasks. The synchronous execution model exposes the overhead of gradient communication among a large number of workers in a distributed training system.

To this end, we propose a new SGD algorithm with delayed averaging, namely DaSGD, which can fully parallelize SGD and forward/backward propagations to hide 100\% of gradient communication. By adjusting the gradient update scheme, this algorithm uses hardware resources more efficiently and reduces the reliance on high-throughput inter-connects. The theoretical analysis and experimental results conducted in this paper both show its convergence rate of $ O (1 / \sqrt {K})$ stays the same as Mini-batch SGD. A analytical model shows that it enables linear performance scalability with the cluster size.

****

Model-Invariant State Abstractions for Model-Based Reinforcement Learning

Manan Tomar,Amy Zhang,Roberto Calandra,Matthew E. Taylor,Joelle Pineau

Accuracy and generalization of dynamics models is key to the success of model-based reinforcement learning (MBRL). As the complexity of tasks increases, learning accurate dynamics models becomes increasingly sample inefficient. However, many complex tasks also exhibit sparsity in dynamics, i.e., actions have only a local effect on the system dynamics. In this paper, we exploit this property with a causal invariance perspective in the single-task setting, introducing a new type of state abstraction called \textit{model-invariance}. Unlike previous forms of state abstractions, a model-invariance state abstraction leverages causal sparsity over state variables. This allows for compositional generalization to unseen states, something that non-factored forms of state abstractions cannot do. We prove that an optimal policy can be learned over this model-invariance state abstraction and show improved generalization in a simple toy domain. Next, we propose a practical method to approximately learn a model-invariant representation for complex domains and validate our approach by showing improved modelling performance over standard maximum likelihood approaches on challenging tasks, such as the MuJoCo-based Humanoid. Finally, within the MBRL setting we show strong performance gains with respect to sample efficiency across a host of continuous control tasks.

****

PARL: Enhancing Diversity of Ensemble Networks to Resist Adversarial Attacks via Pairwise Adversarially Robust Loss Function

Manaar Alam,Shubhajit Datta,Debdeep Mukhopadhyay,Arijit Mondal,Partha Pratim Chakrabarti

The security of Deep Learning classifiers is a critical field of study because of the existence of adversarial attacks. Such attacks usually rely on the principle of transferability, where an adversarial example crafted on a surrogate classifier tends to mislead the target classifier trained on the same dataset even if both classifiers have quite different architecture. Ensemble methods against adversarial attacks demonstrate that an adversarial example is less likely to mislead multiple classifiers in an ensemble having diverse decision boundaries. However, recent ensemble methods have either been shown to be vulnerable to stronger adversaries or shown to lack an end-to-end evaluation. This paper attempts to develop a new ensemble methodology that constructs multiple diverse classifiers using a Pairwise Adversarially Robust Loss (PARL) function during the training procedure. PARL utilizes gradients of each layer with respect to input in every classifier within the ensemble simultaneously. The proposed training procedure enables PARL to achieve higher robustness with high clean example accuracy against black-box transfer attacks compared to the previous ensemble methods. We also evaluate the robustness in the presence of white-box attacks, where adversarial examples are crafted on the target classifier. We present extensive experiments using standard image classification datasets like CIFAR-10 and CIFAR-100 trained using standard ResNet20 classifier against state-of-the-art adversarial attacks to demonstrate the robustness of the proposed ensemble methodology.

****

Recursive Construction of Stable Assemblies of Recurrent Neural Networks

Leo Kozachkov,Michaela M Ennis,Jean-Jacques Slotine

Advanced applications of modern machine learning will likely involve combinations of trained networks, as are already used in spectacular systems such as DeepMind's AlphaGo. Recursively building such combinations in an effective and stable fashion while also allowing for continual refinement of the individual networks - as nature does for biological networks - will require new analysis tools. This paper takes a step in this direction by establishing contraction properties of broad classes of nonlinear recurrent networks and neural ODEs, and showing how these quantified properties allow in turn to recursively construct stable networks of networks in a systematic fashion. The results can also be used to stably combine recurrent networks and physical systems with quantified contraction properties. Similarly, they may be applied to modular computational models of cognition. We perform experiments with these combined networks on benchmark sequential tasks (e.g permuted sequential MNIST) to demonstrate their capacity for processing information across a long timescale in a provably stable manner.

****

How Faithful is your Synthetic Data? Sample-level Metrics for Evaluating and Auditing Generative Models

Ahmed Alaa,Boris van Breugel,Evgeny Saveliev,Mihaela van der Schaar

Devising domain- and model-agnostic evaluation metrics for generative models is an important and as yet unresolved problem. Most existing metrics, which were tailored solely to the image synthesis setup, exhibit a limited capacity for diagnosing the modes of failure of generative models across broader application domains. In this paper, we introduce a 3-dimensional evaluation metric, (α-Precision, β-Recall, Authenticity), that characterizes the fidelity, diversity and generalization performance of any generative model in a domain-agnostic fashion. Our metric unifies statistical divergence measures with precision-recall analysis, enabling sample-level and distribution-level diagnoses of model fidelity and diversity. We introduce generalization as an additional dimension for model performance that quantifies the extent to which a model copies training data---a crucial performance indicator when modeling sensitive data with requirements on privacy. The three metric components correspond to (interpretable) probabilistic quantities, and are estimated via sample-level binary classification. The

sample-level nature of our metric inspires a novel use case which we call model auditing, wherein we judge the quality of individual samples generated by a (black-box) model, discarding low-quality samples and hence improving the overall model performance in a post-hoc manner.
****

Learning to Solve Multi-Robot Task Allocation with a Covariant-Attention based Neural Architecture

Steve Paul,Payam Ghassemi,Souma Chowdhury

This paper demonstrates how time-constrained multi-robot task allocation (MRTA) problems can be modeled as a Markov Decision Process (MDP) over graphs, such that approximate solutions can be modeled as a policy using Reinforcement Learning (RL) methods.

Inspired by emerging approaches for learning to solve related combinatorial optimization (CO) problems such as multi-traveling salesman (mTSP) problems, a graph neural architecture is conceived in this paper to model the MRTA policy. The generalizability and scalability needs of the complex CO problem presented by MRTA are addressed by innovatively using the concept of Covariant Compositional Networks (CCN) to learn the local structures of graphs. The resulting learning architecture is called Covariant Attention-based Mechanism or CAM, which comprises: 1) an encoder: CCN-based embedding model to represent the task space as learnable feature vectors, 2) a decoder: an attention-based model to facilitate sequential decision outputs, and 3) context: to represent the state of the mission and the robots. To learn the feature vectors, a policy-gradient method is used. The CAM architecture is found to generally outperform a state-of-the-art encoder-decoder method that is purely based on Multi-head Attention (MHA) mechanism in terms of task completion and cost function, when applied to a class of MRTA problems with time deadlines, robot ferry range constraints, and multi-tour allowance. CAM also demonstrated significantly better scalability in terms of cost function over unseen scenarios with larger task/robot spaces than those used for training. Lastly, evidence regarding the unique potential of learning-based approaches in delivering highly time-efficient solutions is provided for a benchmark vehicle routing problem -- where solutions are achieved 100-1000 times faster compared to a non-learning baseline, and for a benchmark MRTA problem with time and capacity constraints -- where solutions for larger problems are achieved 10 times faster compared to non-learning baselines.
****

Exploring General Intelligence of Program Analysis for Multiple Tasks

Yixin Guo,Pengcheng Li,Yingwei Luo,Xiaolin Wang,Zhenlin Wang

Artificial intelligence are gaining more attractions for program analysis and semantic understanding. Nowadays, the prevalent program embedding techniques usually target at one single task, for example detection of binary similarity, program classification, program comment auto-complement, etc, due to the ever-growing program complexities and scale. To this end, we explore a generic program embedding approach that aim at solving multiple program analysis tasks. We design models to extract features of a program, represent the program as an embedding, and use this embedding to solve various analysis tasks. Since different tasks require not only access to the features of the source code, but also are highly relevant to its compilation process, traditional source code or AST-based embedding approaches are no longer applicable. Therefore, we propose a new program embedding approach that constructs a program representation based on the assembly code and simultaneously exploits the rich graph structure information present in the program. We tested our model on two tasks, program classification and binary similarity detection, and obtained accuracy of
80.35% and 45.16%, respectively.
****

Gradual Domain Adaptation in the Wild: When Intermediate Distributions are Absent

Samira Abnar,Rianne van den Berg,Golnaz Ghiasi,Mostafa Dehghani,Nal Kalchbrenner,Hanie Sedghi

We focus on the problem of domain adaptation when the goal is shifting the model towards the target distribution, rather than learning domain invariant representations. It is shown that under the following two assumptions: (a) access to samples from intermediate distributions, and (b) samples being annotated with the amount of change from the source distribution; self-training can be successfully applied on gradually shifted samples to adapt the model toward the target distribution. We hypothesize having (a) is enough to enable iterative self-training to slowly adapt the model to the target distribution, by making use of an implicit curriculum. In the case where (a) does not hold, we observe that iterative self-training falls short. We propose GIFT (Gradual Interpolation of Features toward Target), a method that creates virtual samples from intermediate distributions by interpolating representations of examples from source and target domains.

Our analysis of various synthetic distribution shifts shows that in the presence of (a) iterative self-training naturally forms a curriculum of samples which helps the model to adapt better to the target domain. Furthermore, we show that when (a) does not hold, more iterations hurt the performance of self-training, and in these settings GIFT is advantageous. Additionally, we evaluate self-training, iterative self-training and GIFT on two benchmarks with different types of natural distribution shifts and show that when applied on top of other domain adaptation methods, GIFT improves the performance of the model on the target dataset.

****

Locality-Based Mini Batching for Graph Neural Networks

Johannes Klicpera,Chendi Qian,Stephan Günnemann

Training graph neural networks on large graphs is challenging since there is no clear way of how to extract mini batches from connected data. To solve this, previous methods have primarily relied on sampling. While this often leads to good convergence, it introduces significant overhead and requires expensive random data accesses. In this work we propose locality-based mini batching (LBMB), which circumvents sampling by using fixed mini batches based on node locality. LBMB first partitions the training/validation nodes into batches, and then selects the most important auxiliary nodes for each batch using local clustering. Thanks to precomputed batches and consecutive memory accesses, LBMB accelerates training by up to 20x per epoch compared to previous methods, and thus provides significantly better convergence per runtime. Moreover, it accelerates inference by up to 100x, at little to no cost of accuracy.

****

Learning Time-dependent PDE Solver using Message Passing Graph Neural Networks

Pourya Pilva,Ahmad Zareei

One of the main challenges in solving time-dependent partial differential equations is to develop computationally efficient solvers that are accurate and stable. Here, we introduce a general graph neural network approach to finding efficient PDE solvers through learning using message-passing models. We first introduce domain invariant features for PDE-data inspired by classical PDE solvers for an efficient physical representation. Next, we use graphs to represent PDE-data on an unstructured mesh and show that message passing graph neural networks (MPGNN) can parameterize governing equations, and as a result, efficiently learn accurate solver schemes for linear/nonlinear PDEs. We further show that the solvers are independent of the initial training geometry and can solve the same PDE on more complex domains. Lastly, we show that a recurrent graph neural network approach can find a temporal sequence of solutions to a PDE.

****

Asynchronous Multi-Agent Actor-Critic with Macro-Actions

Yuchen Xiao,Weihao Tan,Christopher Amato

Many realistic multi-agent problems naturally require agents to be capable of performing asynchronously without waiting for other agents to terminate (e.g., multi-robot domains). Such problems can be modeled as Macro-Action Decentralized Partially Observable Markov Decision Processes (MacDec-POMDPs). Current policy gradient methods are not applicable to the asynchronous actions in MacDec-POMDPs, as these methods 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 (macro-action-based) policies in three standard training paradigms: decentralized learning, centralized learning, and centralized training for decentralized execution. Empirical results in various domains show high-quality solutions can be learned for large domains when using our methods.
****

Proof Artifact Co-Training for Theorem Proving with Language Models
Jesse Michael Han,Jason Rute,Yuhuai Wu,Edward Ayers,Stanislas Polu
Labeled data for imitation learning of theorem proving in large libraries of formalized mathematics is scarce as such libraries require years of concentrated effort by human specialists to be built. This is particularly challenging when applying large Transformer language models to tactic prediction, because the scaling of performance with respect to model size is quickly disrupted in the data-scarce, easily-overfitted regime. We propose PACT (Proof Artifact Co-Training), a general methodology for extracting abundant self-supervised data from kernel-level proof terms for joint training alongside the usual tactic prediction objective. We apply this methodology to Lean,an interactive proof assistant which hosts some of the most sophisticated formalized mathematics to date. We instrument Lean with a neural theorem prover driven by a Transformer language model and show that PACT improves theorem proving success rate on a held-out suite of test theorems from 32% to 48%.
****

Synthetic Reduced Nearest Neighbor Model for Regression
Pooya Tavallali,Vahid Behzadan,Mukesh Singhal
Nearest neighbor models are among the most established and accurate approaches to machine learning. In this paper, we investigate Synthetic Reduced Nearest Neighbor (SRNN) as a novel approach to regression tasks. Existing prototype nearest neighbor models are initialized by training a k-means model over each class. However, such initialization is only applicable to classification tasks. In this work, we propose a novel initialization and expectation maximization approach for enabling the application of SRNN to regression. The proposed initialization approach is based on applying the k-means algorithm on the target responses of samples to create various clusters of targets. This is proceeded by learning several centroids in the input space for each cluster found over the targets. Essentially, the initialization consists of finding target clusters and running k-means in the space of feature vectors for the corresponding target cluster. The optimization procedure consists of applying an expectation maximization approach similar to the k-means algorithm that optimizes the centroids in the input space. This algorithm is comprised of two steps: (1) The assignment step, where assignments of the samples to each centroid is found and the target response (i.e., prediction) of each centroid is determined; and (2) the update/centroid step, where each centroid is updated such that the loss function of the entire model is minimized. We will show that the centroid step operates over all samples via solving a weighted binary classification. However, the centroid step is NP-hard and no surrogate objective function exists for solving this problem. Therefore, a new surrogate is proposed to approximate the solution for the centroid step. Furthermore, we consider the consistency of the model, and show that the model is consistent under mild assumptions. The bias-variance relationship in this model is also discussed. We report the empirical evaluation of the proposed SRNN regression model in comparison to several state-of-the-art techniques.
****

Mirror Descent Policy Optimization
Manan Tomar,Lior Shani,Yonathan Efroni,Mohammad Ghavamzadeh
Mirror descent (MD), a well-known first-order method in constrained convex optimization, has recently been shown as an important tool to analyze trust-region algorithms in reinforcement learning (RL). However, there remains a considerable gap between such theoretically analyzed algorithms and the ones used in practice. Inspired by this, we propose an efficient RL algorithm, called {\em mirror descent policy optimization} (MDPO). MDPO iteratively updates the policy by {\em approximately} solving a trust-region problem, whose objective function consists of two terms: a linearization of the standard RL objective and a proximity term th

at restricts two consecutive policies to be close to each other. Each update performs this approximation by taking multiple gradient steps on this objective function. We derive {\em on-policy} and {\em off-policy} variants of MDPO, while emphasizing important design choices motivated by the existing theory of MD in RL. We highlight the connections between on-policy MDPO and two popular trust-region RL algorithms: TRPO and PPO, and show that explicitly enforcing the trust-region constraint is in fact {\em not} a necessity for high performance gains in TRPO. We then show how the popular soft actor-critic (SAC) algorithm can be derived by slight modifications of off-policy MDPO. Overall, MDPO is derived from the MD principles, offers a unified approach to viewing a number of popular RL algorithms, and performs better than or on-par with TRPO, PPO, and SAC in a number of continuous and discrete control tasks.

****

MIDI-DDSP: Detailed Control of Musical Performance via Hierarchical Modeling

Yusong Wu,Ethan Manilow,Yi Deng,Rigel Swavely,Kyle Kastner,Tim Cooijmans,Aaron Courville,Cheng-Zhi Anna Huang,Jesse Engel

Musical expression requires control of both what notes that are played, and how they are performed. Conventional audio synthesizers provide detailed expressive controls, but at the cost of realism. Black-box neural audio synthesis and concatenative samplers can produce realistic audio, but have few mechanisms for control. In this work, we introduce MIDI-DDSP a hierarchical model of musical instruments that enables both realistic neural audio synthesis and detailed user control. Starting from interpretable Differentiable Digital Signal Processing (DDSP) synthesis parameters, we infer musical notes and high-level properties of their expressive performance (such as timbre, vibrato, dynamics, and articulation). This creates a 3-level hierarchy (notes, performance, synthesis) that affords individuals the option to intervene at each level, or utilize trained priors (performance given notes, synthesis given performance) for creative assistance. Through quantitative experiments and listening tests, we demonstrate that this hierarchy can reconstruct high-fidelity audio, accurately predict performance attributes for a note sequence, independently manipulate the attributes of a given performance, and as a complete system, generate realistic audio from a novel note sequence. By utilizing an interpretable hierarchy, with multiple levels of granularity, MIDI-DDSP opens the door to assistive tools to empower individuals across a diverse range of musical experience.

****

If your data distribution shifts, use self-learning

Evgenia Rusak,Steffen Schneider,George Pachitariu,Luisa Eck,Peter Vincent Gehler,Oliver Bringmann,Wieland Brendel,Matthias Bethge

In this paper, we demonstrate that self-learning techniques like entropy minimization or pseudo-labeling are simple, yet effective techniques for increasing test performance under domain shifts. Our results show that self-learning consistently increases performance under distribution shifts, irrespective of the model architecture, the pre-training technique or the type of distribution shift. At the same time, self-learning is simple to use in practice because it does not require knowledge or access to the original training data or scheme, is robust to hyperparameter choices, is straight-forward to implement and requires only a few training epochs. This makes self-learning techniques highly attractive for any practitioner who applies machine learning algorithms in the real world. We present state-of-the art adaptation results on CIFAR10-C (8.5% error), ImageNet-C (22.0% mCE), ImageNet-R (17.4% error) and ImageNet-A (14.8% error), theoretically study the dynamics of self-supervised adaptation methods and propose a new classification dataset (ImageNet-D) which is challenging even with adaptation.

****

Exploring the Limits of Large Scale Pre-training

Samira Abnar,Mostafa Dehghani,Behnam Neyshabur,Hanie Sedghi

Recent developments in large-scale machine learning suggest that by scaling up data, model size and training time properly, one might observe that improvements in pre-training would transfer favorably to most downstream tasks. In this work we systematically study this phenomena and establish that, as we increase the upstream accuracy, performance of downstream tasks \emph{saturates}. In particular, we investigate more than 4800 experiments on Vision Transformers, MLP-Mixers and ResNets with number of parameters ranging from ten million to ten billion, trained on the largest scale of available image data (JFT, ImageNet21K) and evaluated on more than 20 downstream image recognition tasks. We propose a model for downstream performance that reflects the saturation phenomena and captures the nonlinear relationship in performance of upstream and downstream tasks. Delving deeper to understand the reasons that give rise to these phenomena, we show that the observed saturation behavior is closely related to the way that representations evolve through the layers of the models. We showcase an even more extreme scenario where performance on upstream and downstream are at odds with each other. That is, in order to have a better downstream performance, we need to hurt upstream accuracy.

****

Taking ROCKET on an efficiency mission: A distributed solution for fast and accurate multivariate time series classification

Leonardos Pantiskas,Kees Verstoep,Mark Hoogendoorn,Henri Bal

Nowadays, with the rising number of sensors in sectors such as healthcare and industry, the problem of multivariate time series classification (MTSC) is getting increasingly relevant and is a prime target for machine and deep learning solutions. Their expanding adoption in real-world environments is causing a shift in focus from the pursuit of ever higher prediction accuracy with complex models towards practical, deployable solutions that balance accuracy and parameters such as prediction speed. An MTSC solution that has attracted attention recently is ROCKET, based on random convolutional kernels, both because of its very fast training process and its state-of-the-art accuracy. However, the large number of features it utilizes may be detrimental to inference time. Examining its theoretical background and limitations enables us to address potential drawbacks and present LightWaveS: a distributed solution for accurate MTSC, which is fast both during training and inference. Specifically, utilizing a wavelet scattering transformation of the time series and distributed feature selection, we manage to create a solution which employs just 2,5% of the ROCKET features, while achieving accuracy comparable to recent deep learning solutions. LightWaveS also scales well with more nodes and large numbers of channels. In addition, it can give interpretability into the nature of an MTSC problem and allows for tuning based on expert opinion. We present three versions of our algorithm and their results on training time, accuracy, inference speedup and scalability. We show that we achieve speedup ranging from 8x to 30x compared to ROCKET during inference on an edge device, on datasets with comparable accuracy.

****

Universal Approximation Under Constraints is Possible with Transformers

Anastasis Kratsios,Behnoosh Zamanlooy,Tianlin Liu,Ivan Dokmani■

Many practical problems need the output of a machine learning model to satisfy a set of constraints, K. Nevertheless, there is no known guarantee that classical neural network architectures can exactly encode constraints while simultaneously achieving universality. We provide a quantitative constrained universal approximation theorem which guarantees that for any non-convex compact set K and any continuous function $f:\mathbb{R}^n\rightarrow K$, there is a probabilistic transformer $\hat{F}$ whose randomized outputs all lie in K and whose expected output uniformly approximates f. Our second main result is a ``deep neural version'' of Berge's Maximum Theorem (1963). The result guarantees that given an objective function L, a constraint set K, and a family of soft constraint sets, there is a probabilistic transformer $\hat{F}$ that approximately minimizes L and whose outputs belong to K; moreover, $\hat{F}$ approximately satisfies the soft constraints. Our results imply the first universal approximation theorem for classical transformers with exact convex constraint satisfaction. They also yield that a chart-free universal approximation theorem for Riemannian manifold-valued functions subject to suitable geodesically convex constraints.

****

Pretrained Language Models are Symbolic Mathematics Solvers too!

Kimia Noorbakhsh,Modar Sulaiman,Mahdi Sharifi,KALLOL ROY,Pooyan Jamshidi

Solving symbolic mathematics has always been of in the arena of human ingenuity that needs compositional reasoning and recurrence. However, recent studies have shown that large scale language models such as transformers are universal and surprisingly can be trained as a sequence-to-sequence task to solve complex mathematical equations. These large transformer models need humongous amounts of training data to generalize to unseen symbolic mathematics problems. In this paper, we present a sample efficient way of solving the symbolic tasks by first pretraining the transformer model with language translation and then fine-tuning the pretrained transformer model to solve the downstream task of symbolic mathematics . We achieve comparable accuracy on the integration task with our pretrained model while using around 1.5 orders of magnitude less number of training samples with respect to the state-of-the-art deep learning for symbolic mathematics. The test accuracy on differential equation tasks is considerably lower comparing with integration as they need higher order recursions that are not present in language translations. We pretrain our model with different pairs of language translations. Our results show language bias in solving symbolic mathematics tasks. Finally, we study the robustness of the fine-tuned model on symbolic math tasks against distribution shift, and our approach generalizes better in distribution shift scenarios for the function integration.

****

A Study on Representation Transfer for Few-Shot Learning

Chun-Nam Yu,Yi Xie

Few-shot classification aims to learn to classify new object categories well using only a few labeled examples. Transfering feature representations from other models is a popular approach for solving few-shot classification problems.In this work we perform a systematic study of various feature representations for few-shot classification, including representations learned from MAML, supervised classification, and several common self-supervised tasks. We find that learning from more complex tasks tend to give better representations for few-shot classification, and thus we propose the use of representations learned from multiple tasks for few-shot classification. Coupled with new tricks on feature selection and voting to handle the issue of small sample size, our direct transfer learning method offers performance comparable to state-of-art on several benchmark datasets.

****

Personalized Heterogeneous Federated Learning with Gradient Similarity

Jing Xie,Xiang Yin,Xiyi Zhang,Juan Chen,Quan Wen,Qiang Yang,Xuan Mo

In the conventional federated learning (FL), the local models of multiple clients are trained independently by their privacy data, and the center server generates the shared global model by aggregating local models. However, the global model often fails to adapt to each client due to statistical and systems heterogeneities, such as non-IID data and inconsistencies in clients' hardware and bandwidth. To address these problems, we propose the Subclass Personalized FL (SPFL) algorithm for non-IID data in synchronous FL and the Personalized Leap Gradient Approximation (PLGA) algorithm for the asynchronous FL. In SPFL, the server uses the Softmax Normalized Gradient Similarity (SNGS) to weight the relationship between clients, and sends the personalized global model to each client. In PLGA, the server also applies the SNGS to weight the relationship between client and itself, and uses the first-order Taylor expansion of gradient to approximate the model of the delayed clients. To the best of our knowledge, this is one of the few studies investigating explicitly on personalization in asynchronous FL. The stage strategy of ResNet is further applied to improve the performance of FL. The experimental results show that (1) in synchronous FL, the SPFL algorithm used on non-IID data outperforms the vanilla FedAvg, PerFedAvg, and FedUpdate algorithms, improving the accuracy by $1.81\!\sim\!18.46\%$ on four datasets (CIFAR10 , CIFAR100, MNIST, EMNIST), while still maintaining the state of the art performance on IID data; (2) in asynchronous FL, compared with the vanilla FedAvg, PerFedAvg, and FedAsync algorithms, the PLGA algorithm improves the accuracy by $0.

23\!\sim\!12.63\%$ on the same four datasets of non-IID data.

****

A Loss Curvature Perspective on Training Instabilities of Deep Learning Models
Justin Gilmer,Behrooz Ghorbani,Ankush Garg,Sneha Kudugunta,Behnam Neyshabur,David Cardoze,George Edward Dahl,Zachary Nado,Orhan Firat
In this work, we study the evolution of the loss Hessian across many classification tasks in order to understand the effect the curvature of the loss has on the training dynamics. Whereas prior work has focused on how different learning rates affect the loss Hessian observed during training, we also analyze the effects of model initialization, architectural choices, and common training heuristics such as gradient clipping and learning rate warmup. Our results demonstrate that successful model and hyperparameter choices allow the early optimization trajectory to either avoid---or navigate out of---regions of high curvature and into flatter regions that tolerate a higher learning rate. Our results suggest a unifying perspective on how disparate mitigation strategies for training instability ultimately address the same underlying failure mode of neural network optimization, namely poor conditioning. Inspired by the conditioning perspective, we show that learning rate warmup can improve training stability just as much as batch normalization, layer normalization, MetaInit, GradInit, and Fixup initialization.

****

Second-Order Unsupervised Feature Selection via Knowledge Contrastive Distillation
Han Yue,Jundong Li,Hongfu Liu
Unsupervised feature selection aims to select a subset from the original features that are most useful for the downstream tasks without external guidance information. While most unsupervised feature selection methods focus on ranking features based on the intrinsic properties of data, they do not pay much attention to the relationships between features, which often leads to redundancy among the selected features. In this paper, we propose a two-stage Second-Order unsupervised Feature selection via knowledge contrastive disTillation (SOFT) model that incorporates the second-order covariance matrix with the first-order data matrix for unsupervised feature selection. In the first stage, we learn a sparse attention matrix that can represent second-order relations between features. In the second stage, we build a relational graph based on the learned attention matrix and perform graph segmentation for feature selection. Experimental results on 12 public datasets show that SOFT outperforms classical and recent state-of-the-art methods, which demonstrates the effectiveness of our proposed method.

****

Cross-Domain Imitation Learning via Optimal Transport
Arnaud Fickinger,Samuel Cohen,Stuart Russell,Brandon Amos
Cross-domain imitation learning studies how to leverage expert demonstrations of one agent to train an imitation agent with a different embodiment or morphology. Comparing trajectories and stationary distributions between the expert and imitation agents is challenging because they live on different systems that may not even have the same dimensionality. We propose Gromov-Wasserstein Imitation Learning (GWIL), a method for cross-domain imitation that uses the Gromov-Wasserstein distance to align and compare states between the different spaces of the agents. Our theory formally characterizes the scenarios where GWIL preserves optimality, revealing its possibilities and limitations. We demonstrate the effectiveness of GWIL in non-trivial continuous control domains ranging from simple rigid transformation of the expert domain to arbitrary transformation of the state-action space.

****

Large-Scale Representation Learning on Graphs via Bootstrapping
Shantanu Thakoor,Corentin Tallec,Mohammad Gheshlaghi Azar,Mehdi Azabou,Eva L Dyer,Remi Munos,Petar Veli■kovi■,Michal Valko
Self-supervised learning provides a promising path towards eliminating the need for costly label information in representation learning on graphs. However, to achieve state-of-the-art performance, methods often need large numbers of negative examples and rely on complex augmentations. This can be prohibitively expens

ive, especially for large graphs. To address these challenges, we introduce Bootstrapped Graph Latents (BGRL) - a graph representation learning method that learns by predicting alternative augmentations of the input. BGRL uses only simple augmentations and alleviates the need for contrasting with negative examples, and thus is scalable by design. BGRL outperforms or matches prior methods on several established benchmarks, while achieving a 2-10x reduction in memory costs. Furthermore, we show that BGRL can be scaled up to extremely large graphs with hundreds of millions of nodes in the semi-supervised regime, achieving state-of-the-art performance and improving over supervised baselines where representations are shaped only through label information. In particular, our solution centered on BGRL constituted one of the winning entries to the Open Graph Benchmark -Large Scale Challenge at KDD Cup 2021, on a graph orders of magnitudes larger than all previously available benchmarks, thus demonstrating the scalability and effectiveness of our approach.

****

Robust and Scalable SDE Learning: A Functional Perspective

Scott Alexander Cameron,Tyron Luke Cameron,Arnu Pretorius,Stephen J. Roberts

Stochastic differential equations provide a rich class of flexible generative models, capable of describing a wide range of spatio-temporal processes. A host of recent work looks to learn data-representing SDEs, using neural networks and other flexible function approximators. Despite these advances, learning remains computationally expensive due to the sequential nature of SDE integrators. In this work, we propose an importance-sampling estimator for probabilities of observations of SDEs for the purposes of learning. Crucially, the approach we suggest does not rely on such integrators. The proposed method produces lower-variance gradient estimates compared to algorithms based on SDE integrators and has the added advantage of being embarrassingly parallelizable. This facilitates the effective use of large-scale parallel hardware for massive decreases in computation time.

****

From SCAN to Real Data: Systematic Generalization via Meaningful Learning

Ning Shi,Boxin Wang,Wei Wang,Xiangyu Liu,Rong Zhang,Hui Xue',Xinbing Wang,Zhouhan Lin

Humans can systematically generalize to novel compositions of existing concepts. There have been extensive conjectures into the extent to which neural networks can do the same. Recent arguments supported by evidence on the SCAN dataset claim that neural networks are inherently ineffective in such cognitive capacity. In this paper, we revisit systematic generalization from the perspective of meaningful learning, an exceptional capability of humans to learn new concepts by connecting them with other previously known knowledge. We propose to augment a training dataset in either an inductive or deductive manner to build semantic links between new and old concepts. Our observations on SCAN suggest that, following the meaningful learning principle, modern sequence-to-sequence models, including RNNs, CNNs, and Transformers, can successfully generalize to compositions of new concepts. We further validate our findings on two real-world datasets on semantic parsing and consistent compositional generalization is also observed. Moreover, our experiments demonstrate that both prior knowledge and semantic linking play a key role to achieve systematic generalization. Meanwhile, inductive learning generally works better than deductive learning in our experiments. Finally, we provide an explanation for data augmentation techniques by concluding them into either inductive-based or deductive-based meaningful learning. We hope our findings will encourage excavating existing neural networks' potential in systematic generalization through more advanced learning schemes.

****

Exploiting Redundancy: Separable Group Convolutional Networks on Lie Groups

David M Knigge,David W. Romero,Erik J Bekkers

Group convolutional neural networks (G-CNNs) have been shown to increase parameter efficiency and model accuracy by incorporating geometric inductive biases. In this work, we investigate the properties of representations learned by regular

G-CNNs, and show considerable parameter redundancy in group convolution kernels. This finding motivates further weight-tying by sharing convolution kernels over subgroups. To this end, we introduce convolution kernels that are separable over the subgroup and channel dimensions. In order to obtain equivariance to arbitrary affine Lie groups we provide a continuous parameterisation of separable convolution kernels. We evaluate our approach across several vision datasets, and show that our weight sharing leads to improved performance and computational efficiency. In many settings, separable G-CNNs outperform their non-separable counterpart, while only using a fraction of their training time. In addition, thanks to the increase in computational efficiency, we are able to implement G-CNNs equivariant to the $\mathrm{Sim(2)}$ group; the group of dilations, rotations and translations. $\mathrm{Sim(2)}$-equivariance further improves performance on all tasks considered.
****

Neural Processes with Stochastic Attention: Paying more attention to the context dataset
Mingyu Kim,Kyeong Ryeol Go,Se-Young Yun
Neural processes (NPs) aim to stochastically complete unseen data points based on a given context dataset. NPs essentially leverage a given dataset as a context representation to derive a suitable identifier for a novel task. To improve the prediction accuracy, many variants of NPs have investigated context embedding approaches that generally design novel network architectures and aggregation functions satisfying permutation invariant. In this work, we propose a stochastic attention mechanism for NPs to capture appropriate context information. From the perspective of information theory, we demonstrate that the proposed method encourages context embedding to be differentiated from a target dataset, allowing NPs to consider features in a target dataset and context embedding independently. We observe that the proposed method can appropriately capture context embedding even under noisy data sets and restricted task distributions, where typical NPs suffer from a lack of context embeddings. We empirically show that our approach substantially outperforms conventional NPs in various domains through 1D regression, predator-prey model, and image completion. Moreover, the proposed method is also validated by MovieLens-10k dataset, a real-world problem.
****

Evaluating Disentanglement of Structured Representations
Raphaël Dang-Nhu
We introduce the first metric for evaluating disentanglement at individual hierarchy levels of a structured latent representation. Applied to object-centric generative models, this offers a systematic, unified approach to evaluating (i) object separation between latent slots (ii) disentanglement of object properties inside individual slots (iii) disentanglement of intrinsic and extrinsic object properties. We theoretically show that our framework gives stronger guarantees of selecting a good model than previous disentanglement metrics. Experimentally, we demonstrate that viewing object compositionality as a disentanglement problem addresses several issues with prior visual metrics of object separation. As a core technical component, we present the first representation probing algorithm handling slot permutation invariance.
****

Wavelet-Packet Powered Deepfake Image Detection
Moritz Wolter,Felix Blanke,Charles Tapley Hoyt,Jochen Garcke
As neural networks become able to generate realistic artificial images, they have the potential to improve movies, music, video games and make the internet an even more creative and inspiring place.
Yet, at the same time, the latest technology potentially enables new digital ways to lie. In response, the need for a diverse and reliable method toolbox arises to identify artificial images and other content.
Previous work primarily relies on pixel-space CNN or the Fourier transform. To the best of our knowledge, synthesized fake image analysis and detection methods based on a multi-scale wavelet representation,
which is localized in both space and frequency, have been absent thus far.

This paper proposes to learn a model for the detection of synthetic images based on the wavelet-packet representation of natural and GAN-generated images. We evaluate our method on FFHQ, CelebA, and LSUN source identification problems and find improved or competitive performance. Our forensic classifier has a small network size and can be learned efficiently.

Furthermore, a comparison of the wavelet coefficients from these two sources of images allows an interpretation and identifies significant differences.

****

Geometric Transformers for Protein Interface Contact Prediction

Alex Morehead,Chen Chen,Jianlin Cheng

Computational methods for predicting the interface contacts between proteins come highly sought after for drug discovery as they can significantly advance the accuracy of alternative approaches, such as protein-protein docking, protein function analysis tools, and other computational methods for protein bioinformatics. In this work, we present the Geometric Transformer, a novel geometry-evolving graph transformer for rotation and translation-invariant protein interface contact prediction, packaged within DeepInteract, an end-to-end prediction pipeline. DeepInteract predicts partner-specific protein interface contacts (i.e., inter-protein residue-residue contacts) given the 3D tertiary structures of two proteins as input. In rigorous benchmarks, DeepInteract, on challenging protein complex targets from the 13th and 14th CASP-CAPRI experiments as well as Docking Benchmark 5, achieves 14% and 1.1% top L/5 precision (L: length of a protein unit in a complex), respectively. In doing so, DeepInteract, with the Geometric Transformer as its graph-based backbone, outperforms existing methods for interface contact prediction in addition to other graph-based neural network backbones compatible with DeepInteract, thereby validating the effectiveness of the Geometric Transformer for learning rich relational-geometric features for downstream tasks on 3D protein structures.

****

Decision Tree Algorithms for MDP

Elioth Sanabria,David Yao,Henry Lam

Decision trees are robust modeling tools in machine learning with human-interpretable representations. The curse of dimensionality of Markov Decision Process (MDP) makes exact solution methods computationally intractable in practice for large state-action spaces. In this paper, we show that even for problems with large state space, when the solution policy of the MDP can be represented by a tree-like structure, our proposed algorithm retrieves a tree of the solution policy of the MDP in computationally tractable time. Our algorithm uses a tree growing strategy to incrementally disaggregate the state space solving smaller MDP instances with Linear Programming. These ideas can be extended to experience based RL problems as an alternative to black-box based policies.

****

Diurnal or Nocturnal? Federated Learning of Multi-branch Networks from Periodically Shifting Distributions

Chen Zhu,Zheng Xu,Mingqing Chen,Jakub Kone■ný,Andrew Hard,Tom Goldstein

Federated learning has been deployed to train machine learning models from decentralized client data on mobile devices in practice. The clients available for training are observed to have periodically shifting distributions changing with the time of day, which can cause instability in training and degrade the model performance. In this paper, instead of modeling the distribution shift with a block-cyclic pattern as previous works, we model it with a mixture of distributions that gradually shifts between daytime and nighttime modes, and find this intuitive model to better match the observations in practical federated learning systems.

Furthermore, we propose to jointly train a clustering model and a multi-branch network to allocate lightweight specialized branches to clients from different modes. A temporal prior is used to significantly boost the training performance.

Experiments for image classification on EMNIST and CIFAR datasets, and next word prediction on the Stack Overflow dataset show that the proposed algorithm can counter the effects of the distribution shift and significantly improve the final model performance.

****

Newer is not always better: Rethinking transferability metrics, their peculiarities, stability and performance

Shibal Ibrahim,Natalia Ponomareva,Rahul Mazumder

Fine-tuning of large pre-trained image and language models on small customized datasets has become increasingly popular for improved prediction and efficient use of limited resources. Fine-tuning requires identification of best models to transfer-learn from and quantifying transferability prevents expensive re-training on all of the candidate models/tasks pairs. In this paper, we show that the statistical problems with covariance estimation drive the poor performance of H-score (Bao et al., 2019) — a common baseline for newer metrics — and propose shrinkage-based estimator. This results in up to 80% absolute gain in H-score correlation performance, making it competitive with the state-of-the-art LogME measure by You et al. (2021). Our shrinkage-based H-score is 3-55 times faster to compute compared to LogME. Additionally, we look into a less common setting of target (as opposed to source) task selection. We demonstrate previously overlooked problems in such settings with different number of labels, class-imbalance ratios etc. for some recent metrics e.g., NCE (Tran et al., 2019), LEEP (Nguyen et al., 2020) that resulted in them being misrepresented as leading measures. We propose a correction and recommend measuring correlation performance against relative accuracy in such settings. We also outline the difficulties of comparing feature-dependent metrics, both supervised (e.g. H-score) and unsupervised measures (e.g., Maximum Mean (Long et al., 2015) and Central Moment Discrepancy (Zellinger et al., 2019)), across source models/layers with widely varying feature embedding dimension. We show that dimensionality reduction methods allow for meaningful comparison across models, cheaper computation (6x) and improved correlation performance of some of these measures. We investigate performance of 14 different supervised and unsupervised metrics and demonstrate that even unsupervised metrics can identify the leading models for domain adaptation. We support our findings with ~65,000 (fine-tuning trials) experiments.

****

IGLU: Efficient GCN Training via Lazy Updates

S Deepak Narayanan,Aditya Sinha,Prateek Jain,Purushottam Kar,SUNDARARAJAN SELLAMANICKAM

Training multi-layer Graph Convolution Networks (GCN) using standard SGD techniques scales poorly as each descent step ends up updating node embeddings for a large portion of the graph. Recent attempts to remedy this sub-sample the graph that reduces compute but introduce additional variance and may offer suboptimal performance. This paper develops the IGLU method that caches intermediate computations at various GCN layers thus enabling lazy updates that significantly reduce the compute cost of descent. IGLU introduces bounded bias into the gradients but nevertheless converges to a first-order saddle point under standard assumptions such as objective smoothness. Benchmark experiments show that IGLU offers up to 1.2% better accuracy despite requiring up to 88% less compute.

****

Hierarchical Multimodal Variational Autoencoders

Jannik Wolff,Rahul G Krishnan,Lukas Ruff,Jan Nikolas Morshuis,Tassilo Klein,Shinichi Nakajima,Moin Nabi

Humans find structure in natural phenomena by absorbing stimuli from multiple input sources such as vision, text, and speech. We study the use of deep generative models that generate multimodal data from latent representations. Existing approaches generate samples using a single shared latent variable, sometimes with marginally independent latent variables to capture modality-specific variations. However, there are cases where modality-specific variations depend on the kind of structure shared across modalities. To capture such heterogeneity, we propose a hierarchical multimodal VAE (HMVAE) that represents modality-specific variations using latent variables dependent on a shared top-level variable. Our experiments on the CUB and the Oxford Flower datasets show that the HMVAE can represent multimodal heterogeneity and outperform existing methods in sample generation qu ality and quantitative measures as the held-out log-likelihood.

****

VISCOS Flows: Variational Schur Conditional Sampling with Normalizing Flows

Vincent Moens,Aivar Sootla,Haitham Bou Ammar,Jun Wang

We present a method for conditional sampling for pre-trained normalizing flows when only part of an observation is available. We derive a lower bound to the conditioning variable log-probability using Schur complement properties in the spirit of Gaussian conditional sampling. Our derivation relies on partitioning flow's domain in such a way that the flow restrictions to subdomains remain bijective, which is crucial for the Schur complement application. Simulation from the variational conditional flow then amends to solving an equality constraint. Our contribution is three-fold: a) we provide detailed insights on the choice of variational distributions; b) we discuss how to partition the input space of the flow to preserve bijectivity property; c) we propose a set of methods to optimise the variational distribution. Our numerical results indicate that our sampling method can be successfully applied to invertible residual networks for inference and classification.

****

Scaling Laws for Neural Machine Translation

Behrooz Ghorbani,Orhan Firat,Markus Freitag,Ankur Bapna,Maxim Krikun,Xavier Garcia,Ciprian Chelba,Colin Cherry

We present an empirical study of scaling properties of encoder-decoder Transformer models used in neural machine translation (NMT). We show that cross-entropy loss as a function of model size follows a certain scaling law. Specifically (i) We propose a formula which describes the scaling behavior of cross-entropy loss as a bivariate function of encoder and decoder size, and show that it gives accurate predictions under a variety of scaling approaches and languages; we show that the total number of parameters alone is not sufficient for such purposes. (ii) We observe different power law exponents when scaling the decoder vs scaling the encoder, and provide recommendations for optimal allocation of encoder/decoder capacity based on this observation. (iii) We also report that the scaling behavior of the model is acutely influenced by composition bias of the train/test sets, which we define as any deviation from naturally generated text (either via machine generated or human translated text). We observe that natural text on the target side enjoys scaling, which manifests as successful reduction of the cross-entropy loss. (iv) Finally, we investigate the relationship between the cross-entropy loss and the quality of the generated translations. We find two different behaviors, depending on the nature of the test data. For test sets which were originally translated from target language to source language, both loss and BLEU score improve as model size increases. In contrast, for test sets originally translated from source language to target language, the loss improves, but the BLEU score stops improving after a certain threshold. We release generated text from all models used in this study.

****

Procedural generalization by planning with self-supervised world models

Ankesh Anand,Jacob C Walker,Yazhe Li,Eszter Vértes,Julian Schrittwieser,Sherjil Ozair,Theophane Weber,Jessica B Hamrick

One of the key promises of model-based reinforcement learning is the ability to generalize using an internal model of the world to make predictions in novel environments and tasks. However, the generalization ability of model-based agents is not well understood because existing work has focused on model-free agents when benchmarking generalization. Here, we explicitly measure the generalization ability of model-based agents in comparison to their model-free counterparts. We focus our analysis on MuZero (Schrittwieser et al., 2020), a powerful model-based agent, and evaluate its performance on both procedural and task generalization. We identify three factors of procedural generalization---planning, self-supervised representation learning, and procedural data diversity---and show that by combining these techniques, we achieve state-of-the art generalization performance and data efficiency on Procgen (Cobbe et al., 2019). However, we find that these factors do not always provide the same benefits for the task generalization be nchmarks in Meta-World (Yu et al., 2019), indicating that transfer remains a challenge and may require different approaches than procedural generalization. Overall, we suggest that building generalizable agents requires moving beyond the single-task, model-free paradigm and towards self-supervised model-based agents that are trained in rich, procedural, multi-task environments.

****

Top-N: Equivariant Set and Graph Generation without Exchangeability

Clement Vignac,Pascal Frossard

This work addresses one-shot set and graph generation, and, more specifically, the parametrization of probabilistic decoders that map a vector-shaped prior to a distribution over sets or graphs. Sets and graphs are most commonly generated by first sampling points i.i.d. from a normal distribution, and then processing these points along with the prior vector using Transformer layers or Graph Neural Networks.

This architecture is designed to generate exchangeable distributions, i.e., all permutations of the generated outputs are equally likely. We however show that it only optimizes a proxy to the evidence lower bound, which makes it hard to train. We then study equivariance in generative settings and show that non-exchangeable methods can still achieve permutation equivariance. Using this result, we introduce Top-n creation, a differentiable generation mechanism that uses the latent vector to select the most relevant points from a trainable reference set. Top-n can replace i.i.d. generation in any Variational Autoencoder or Generative Adversarial Network. Experimentally, our method outperforms i.i.d. generation by 15% at SetMNIST reconstruction, by 33% at object detection on CLEVR, generates sets that are 74% closer to the true distribution on a synthetic molecule-like dataset, and generates more valid molecules on QM9.

****

On Multi-objective Policy Optimization as a Tool for Reinforcement Learning: Case Studies in Offline RL and Finetuning

Abbas Abdolmaleki,Sandy Huang,Giulia Vezzani,Bobak Shahriari,Jost Tobias Springenberg,Shruti Mishra,Dhruva Tirumala,Arunkumar Byravan,Konstantinos Bousmalis,András György,Csaba Szepesvari,raia hadsell,Nicolas Heess,Martin Riedmiller

Many advances that have improved the robustness and efficiency of deep reinforcement learning (RL) algorithms can, in one way or another, be understood as introducing additional objectives or constraints in the policy optimization step. This includes ideas as far ranging as exploration bonuses, entropy regularization, and regularization toward teachers or data priors. Often, the task reward and auxiliary objectives are in conflict, and in this paper we argue that this makes it natural to treat these cases as instances of multi-objective (MO) optimization problems. We demonstrate how this perspective allows us to develop novel and more effective RL algorithms. In particular, we focus on offline RL and finetuning as case studies, and show that existing approaches can be understood as MO algorithms relying on linear scalarization. We hypothesize that replacing linear scalarization with a better algorithm can improve performance. We introduce Distillation of a Mixture of Experts (DiME), a new MORL algorithm that outperforms linear scalarization and can be applied to these non-standard MO problems. We demonstrate that for offline RL, DiME leads to a simple new algorithm that outperforms state-of-the-art. For finetuning, we derive new algorithms that learn to outperform the teacher policy.

****

Adaptive Pseudo-labeling for Quantum Calculations

Kexin Huang,Vishnu Sresht,Brajesh Rai,Mykola Bordyuh

Machine learning models have recently shown promise in predicting molecular quantum chemical properties. However, the path to real-life adoption requires (1) learning under low-resource constraint and (2) out-of-distribution generalization to unseen, structurally diverse molecules. We observe that these two challenges can be alleviated via abundant labels, which are often not the case in quantum chemistry. We hypothesize that pseudo-labeling on vast array of unlabeled molecules can serve as gold-label proxies to greatly expand the training labeled dataset. The challenge in pseudo-labeling is to prevent the bad pseudo-labels from bia sing the model. We develop a simple and effective strategy Pseudo that can assign pseudo-labels, detect bad pseud-labels through evidential uncertainty, and then prevent them from biasing the model using adaptive weighting. Empirically, Pseudo improves quantum calculations accuracy across full data, low data and out-of-distribution settings.

****

Constraint-based graph network simulator

Yulia Rubanova,Alvaro Sanchez-Gonzalez,Tobias Pfaff,Peter Battaglia

In the rapidly advancing area of learned physical simulators, nearly all methods train a forward model that directly predicts future states from input states. However, many traditional simulation engines use a constraint-based approach instead of direct prediction. Here we present a framework for constraint-based learned simulation, where a scalar constraint function is implemented as a trainable function approximator, and future predictions are computed as the solutions to a constraint satisfaction problem. We implement our method using a graph neural network as the constraint function and gradient descent as the constraint solver. The architecture can be trained by standard backpropagation. We test the model on a variety of challenging physical domains, including simulated ropes, bouncing balls, colliding irregular shapes and splashing fluids. Our model achieves better or comparable performance to top learned simulators. A key advantage of our model is the ability to generalize to more solver iterations at test time to improve the simulation accuracy. We also show how hand-designed constraints can be added at test time to satisfy objectives which were not present in the training data, which is not possible with forward approaches. Our constraint-based framework is applicable to any setting in which forward learned simulators are used, and more generally demonstrates key ways that learned models can leverage popular methods in numerical methods.

****

Graph Similarities and Dual Approach for Sequential Text-to-Image Retrieval

Keonwoo Kim,Sihyeon Jo,Seong-Woo Kim

Sequential text-to-image retrieval, a.k.a. Story-to-images task, requires semantic alignment with a given story and maintaining global coherence in drawn image sequence simultaneously. Most of the previous works have only focused on modeling how to follow the content of a given story faithfully. This kind of overfitting tendency hinders matching structural similarity between images, causing an inconsistency in global visual information such as backgrounds. To handle this imbalanced problem, we propose a novel image sequence retrieval framework that utilizes scene graph similarities of the images and a dual learning scheme. Scene graph describes high-level information of visual groundings and adjacency relations of the key entities in a visual scene. In our proposed retriever, the graph encoding head learns to maximize graph embedding similarities among sampled images, giving a strong signal that forces the retriever to also consider morphological relevance with previously sampled images. We set a video captioning as a dual learning task that reconstructs the input story from the sampled image sequence. This inverse mapping gives informative feedback for our proposed retrieval system to maintain global contextual information of a given story. We also suggest a new contextual sentence encoding architecture to embed a sentence in consideration of the surrounding context. Through extensive experiments, Our proposed framework shows better qualitative and quantitative performance with Visual Storytelling benchmark compared to conventional story-to-image models.

****

The Spectral Bias of Polynomial Neural Networks

Moulik Choraria,Leello Tadesse Dadi,Grigorios Chrysos,Julien Mairal,Volkan Cevher

Polynomial neural networks (PNNs) have been recently shown to be particularly effective at image generation and face recognition, where high-frequency information is critical. Previous studies have revealed that neural networks demonstrate a $\text{\it{spectral bias}}$ towards low-frequency functions, which yields faster learning of low-frequency components during training. Inspired by such studies, we conduct a spectral analysis of the Neural Tangent Kernel (NTK) of PNNs. We

find that the Π-Net family, i.e., a recently proposed parametrization of PNNs, speeds up the learning of the higher frequencies.
We verify the theoretical bias through extensive experiments. We expect our analysis to provide novel insights into designing architectures and learning frameworks by incorporating multiplicative interactions via polynomials.

****

Invariant Causal Representation Learning for Out-of-Distribution Generalization
Chaochao Lu,Yuhuai Wu,José Miguel Hernández-Lobato,Bernhard Schölkopf
Due to spurious correlations, machine learning systems often fail to generalize to environments whose distributions differ from the ones used at training time. Prior work addressing this, either explicitly or implicitly, attempted to find a data representation that has an invariant relationship with the target. This is done by leveraging a diverse set of training environments to reduce the effect of spurious features and build an invariant predictor. However, these methods have generalization guarantees only when both data representation and classifiers come from a linear model class. We propose invariant Causal Representation Learning (iCaRL), an approach that enables out-of-distribution (OOD) generalization in the nonlinear setting (i.e., nonlinear representations and nonlinear classifiers). It builds upon a practical and general assumption: the prior over the data representation (i.e., a set of latent variables encoding the data) given the target and the environment belongs to general exponential family distributions, i.e., a more flexible conditionally non-factorized prior that can actually capture complicated dependences between the latent variables. Based on this, we show that it is possible to identify the data representation up to simple transformations. We also show that all direct causes of the target can be fully discovered, which further enables us to obtain generalization guarantees in the nonlinear setting. Experiments on both synthetic and real-world datasets demonstrate that our approach outperforms a variety of baseline methods.

****

Reasoning-Modulated Representations
Petar Veli■kovi■,Matko Bošnjak,Thomas Kipf,Alexander Lerchner,raia hadsell,Razvan Pascanu,Charles Blundell
Neural networks leverage robust internal representations in order to generalise. Learning them is difficult, and often requires a large training set that covers the data distribution densely. We study a common setting where our task is not purely opaque. Indeed, very often we may have access to information about the underlying system (e.g. that observations must obey certain laws of physics) that any "tabula rasa" neural network would need to re-learn from scratch, penalising data efficiency. We incorporate this information into a pre-trained reasoning module, and investigate its role in shaping the discovered representations in diverse self-supervised learning settings from pixels. Our approach paves the way for a new class of data-efficient representation learning.

****

AdaRL: What, Where, and How to Adapt in Transfer Reinforcement Learning
Biwei Huang,Fan Feng,Chaochao Lu,Sara Magliacane,Kun Zhang
One practical challenge in reinforcement learning (RL) is how to make quick adaptations when faced with new environments. In this paper, we propose a principled framework for adaptive RL, called AdaRL, that adapts reliably and efficiently to changes across domains with a few samples from the target domain, even in partially observable environments. Specifically, we leverage a parsimonious graphical representation that characterizes structural relationships over variables in the RL system. Such graphical representations provide a compact way to encode what and where the changes across domains are, and furthermore inform us with a minimal set of changes that one has to consider for the purpose of policy adaptation. We show that by explicitly leveraging this compact representation to encode changes, we can efficiently adapt the policy to the target domain, in which only a few samples are needed and further policy optimization is avoided. We illustrate the efficacy of AdaRL through a series of experiments that vary factors in the observation, transition and reward functions for Cartpole and Atari games.

****

Interpretable Semantic Role Relation Table for Supporting Facts Recognition of Reading Comprehension
YanQing Bai,Zhichang Zhang,HaoYuan Chen,Xiaohui Qin,Yanglong Qiu
The current Machine Reading Comprehension (MRC) model has poor interpretability.

Interpretable semantic features can enhancethe interpretability of the model.
Semantic role labeling (SRL) captures predicate-argument relations, such as "who did what to whom," which
are critical to comprehension and interpretation.
To enhance the interpretability of the model, we propose the semantic role relation table,
which represents the semantic relation of the sentence itself and the semantic relations among sentences.
We use the name of entities to integrate into the semantic role relation table to establish the semantic relation between sentences.

This paper makes the first attempt to utilize contextual semantic's explicit relation to the recognition supporting fact of reading
comprehension.
We have established nine semantic relationtables between target sentence, question, and article.
Then we take each semantic relationship table's overall semantic role relevance and each
semantic role relevance as important judgment information.
Detailed experiments on HotpotQA, a challenging multi-hop MRC data set, our method achieves better performance.
With few training data sets, the model performance is still stable.
****

Q-learning for real time control of heterogeneous microagent collectives

Ana Rubio Denniss,Laia Freixas Mateu,Thomas Gorochowski,Sabine Hauert

The effective control of microscopic collectives has many promising applications, from environmental remediation to targeted drug delivery. A key challenge is understanding how to control these agents given their limited programmability, and in many cases heterogeneous dynamics. The ability to learn control strategies in real time could allow for the application of robotics solutions to drive collective behaviours towards desired outcomes. Here, we demonstrate Q-learning on the closed-loop Dynamic Optical Micro-Environment (DOME) platform to control the motion of light-responsive Volvox agents. The results show that Q-learning is efficient in autonomously learning how to reduce the speed of agents on an individual basis.
****

Reinforcement Learning with Efficient Active Feature Acquisition

Haiyan Yin,Yingzhen Li,Sinno Pan,Cheng Zhang,Sebastian Tschiatschek

Solving real-life sequential decision making problems under partial observability involves an exploration-exploitation problem. To be successful, an agent needs to efficiently gather valuable information about the state of the world for making rewarding decisions. However, in real-life, acquiring valuable information is often highly costly, e.g., in the medical domain, information acquisition might correspond to performing a medical test on a patient. Thus it poses a significant challenge for the agent to learn optimal task policy while efficiently reducing the cost for information acquisition. In this paper, we introduce a model-based framework to solve such exploration-exploitation problem during its execution. Key to the success is a sequential variational auto-encoder which could learn high-quality representations over the partially observed/missing features, where such representation learning serves as a prime factor to drive efficient policy training under the cost-sensitive setting. We demonstrate our proposed method could significantly outperform conventional approaches in a control domain as well as using a medical simulator.
****

Direct Evolutionary Optimization of Variational Autoencoders With Binary Latents

Enrico Guiraud,Jakob Drefs,Filippos S Panagiotou,Jorg Lucke

Many types of data are generated at least partly by discrete causes that are sparsely
active. To model such data, we here investigate a deep generative model in the
form of a variational autoencoder (VAE) which can learn a sparse, binary code
for its latents. Because of the latents’ discrete nature, standard VAE training is
not possible. The goal of previous approaches has therefore been to amend (i.e.,
typically anneal) discrete priors in order to train discrete VAEs analogously to
conventional ones. Here, we divert much more strongly from conventional VAE
training: We ask if it is also possible to keep the discrete nature of the latents
fully intact by applying a direct, discrete optimization for the encoding model. In
doing so, we (1) sidestep standard VAE mechanisms such as sampling approximation, reparameterization trick and amortization, and (2) observe a much sparser
encoding compared to autoencoders that use annealed discrete latents. Direct opt

imization of VAEs is enabled by an evolutionary algorithm in conjunction with truncated posteriors as variational distributions, i.e. by a combination of methods which is here for the first time applied to a deep model. We first show how the discrete variational method (A) ties into gradient ascent for network weights, and how it (B) uses the decoder network to select binary latent states for training. Sparse codes have prominently been applied to image patches, where latents encode edge-like structure. For our VAEs, we maintain this prototypical application domain and observe the emergence of much sparser codes compared to more conventional VAEs. To allow for a broad comparison to other approaches, the emerging encoding was then evaluated on denoising and inpainting tasks, which are canonically benchmarks for image patch models. For datasets with many, large images of single objects (ImageNet, CIFAR etc) deep generative models with dense codes seem preferable. For image patches, however, we observed advantages of sparse codes that give rise to state-of-the-art performance in ‘zero-shot’ denoising and inpainting benchmarks. Sparse codes can consequently make VAEs competitive on tasks where they have previously been outperformed by non-generative approaches.

****

Automatic prior selection for meta Bayesian optimization with a case study on tuning deep neural network optimizers

Zi Wang,George Edward Dahl,Kevin Swersky,Chansoo Lee,Zelda E Mariet,Zachary Nado,Justin Gilmer,Jasper Snoek,Zoubin Ghahramani

The performance of deep neural networks can be highly sensitive to the choice of a variety of meta-parameters, such as optimizer parameters and model hyperparameters. Tuning these well, however, often requires extensive and costly experimentation. Bayesian optimization (BO) is a principled approach to solve such expensive hyperparameter tuning problems efficiently. Key to the performance of BO is specifying and refining a distribution over functions, which is used to reason about the optima of the underlying function being optimized. In this work, we consider the scenario where we have data from similar functions that allows us to specify a tighter distribution a priori. Specifically, we focus on the common but potentially costly task of tuning optimizer parameters for training neural networks. Building on the meta BO method from Wang et al. (2018), we develop practical improvements that (a) boost its performance by leveraging tuning results on multiple tasks without requiring observations for the same meta-parameter points across all tasks, and (b) retain its regret bound for a special case of our method. As a result, we provide a coherent BO solution for iterative optimization of continuous optimizer parameters. To verify our approach in realistic model training setups, we collected a large multi-task hyperparameter tuning dataset by training tens of thousands of configurations of near-state-of-the-art models on popular image and text datasets, as well as a protein sequence dataset. Our results show that on average, our method is able to locate good hyperparameters at least 3 times more efficiently than the best competing methods.

****

LFPT5: A Unified Framework for Lifelong Few-shot Language Learning Based on Prompt Tuning of T5

Chengwei Qin,Shafiq Joty

Existing approaches to lifelong language learning rely on plenty of labeled data for learning a new task, which is hard to obtain in most real scenarios. Considering that humans can continually learn new tasks from a handful of examples, we expect the models also to be able to generalize well on new few-shot tasks without forgetting the previous ones. In this work, we define this more challenging yet practical problem as Lifelong Few-shot Language Learning (LFLL) and propose a unified framework for it based on prompt tuning of T5. Our framework called LFPT5 takes full advantage of PT's strong few-shot learning ability, and simultaneously trains the model as a task solver and a data generator. Before learning a new domain of the same task type, LFPT5 generates pseudo (labeled) samples of previously learned domains, and later gets trained on those samples to alleviate f

orgetting of previous knowledge as it learns the new domain. In addition, a KL divergence loss is minimized to achieve label consistency between the previous and the current model. While adapting to a new task type, LFPT5 includes and tunes additional prompt embeddings for the new task. With extensive experiments, we demonstrate that LFPT5 can be applied to various different types of tasks and significantly outperform previous methods in different LFLL settings.
****
On Non-Random Missing Labels in Semi-Supervised Learning
Xinting Hu,Yulei Niu,Chunyan Miao,Xian-Sheng Hua,Hanwang Zhang
Semi-Supervised Learning (SSL) is fundamentally a missing label problem, in which the label Missing Not At Random (MNAR) problem is more realistic and challenging, compared to the widely-adopted yet naive Missing Completely At Random assumption where both labeled and unlabeled data share the same class distribution. Different from existing SSL solutions that overlook the role of ''class'' in causing the non-randomness, e.g., users are more likely to label popular classes, we explicitly incorporate ''class'' into SSL. Our method is three-fold: 1) We propose Class-Aware Propensity (CAP) that exploits the unlabeled data to train an improved classifier using the biased labeled data. 2) To encourage rare class training, whose model is low-recall but high-precision that discards too many pseudo-labeled data, we propose Class-Aware Imputation (CAI) that dynamically decreases (or increases) the pseudo-label assignment threshold for rare (or frequent) classes. 3) Overall, we integrate CAP and CAI into a Class-Aware Doubly Robust (CADR) estimator for training an unbiased SSL model. Under various MNAR settings and ablations, our method not only significantly outperforms existing baselines, but also surpasses other label bias removal SSL methods.

****
Equivariant Heterogeneous Graph Networks
Daniel Levy,Siamak Ravanbakhsh
Many real-world datasets include multiple distinct types of entities and relations, and so they are naturally best represented by heterogeneous graphs. However, the most common forms of neural networks operating on graphs either assume that their input graphs are homogeneous, or they convert heterogeneous graphs into homogeneous ones, losing valuable information in the process. Any neural network that acts on graph data should be equivariant or invariant to permutations of nodes, but this is complicated when there are multiple distinct node and edge types. With this as motivation, we design graph neural networks that are composed of linear layers that are maximally expressive while being equivariant only to permutations of nodes within each type. We demonstrate their effectiveness on heterogeneous graph node classification and link prediction benchmarks.
****
On the Evolution of Neuron Communities in a Deep Learning Architecture
Sakib Mostafa,Debajyoti Mondal
Deep learning techniques are increasingly being adopted for classification tasks over the past decade, yet explaining how deep learning architectures can achieve state-of-the-art performance is still an elusive goal. While all the training information is embedded deeply in a trained model, we still do not understand much about its performance by only analyzing the model. This paper examines the neuron activation patterns of deep learning-based classification models and explores whether the models' performances can be explained through neurons' activation behavior. We propose two approaches: one that models neurons' activation behavior as a graph and examines whether the neurons form meaningful communities, and the other examines the predictability of neurons' behavior using entropy. Our comprehensive experimental study reveals that both the community quality and entropy can provide new insights into the deep learning models' performances, thus paves a novel way of explaining deep learning models directly from the neurons' activation pattern
.
****
RAVE: A variational autoencoder for fast and high-quality neural audio synthesis

Antoine Caillon,Philippe Esling

Deep generative models applied to audio have improved by a large margin the state-of-the-art in many speech and music related tasks. However, as raw waveform modelling remains an inherently difficult task, audio generative models are either computationally intensive, rely on low sampling rates, are complicated to control or restrict the nature of possible signals. Among those models, Variational AutoEncoders (VAE) give control over the generation by exposing latent variables, although they usually suffer from low synthesis quality. In this paper, we introduce a Realtime Audio Variational autoEncoder (RAVE) allowing both fast and high-quality audio waveform synthesis. We introduce a novel two-stage training procedure, namely representation learning and adversarial fine-tuning. We show that using a post-training analysis of the latent space allows a direct control between the reconstruction fidelity and the representation compactness. By leveraging a multi-band decomposition of the raw waveform, we show that our model is the first able to generate 48kHz audio signals, while simultaneously running 20 times faster than real-time on a standard laptop CPU. We evaluate synthesis quality using both quantitative and qualitative subjective experiments and show the superiority of our approach compared to existing models. Finally, we present applications of our model for timbre transfer and signal compression. All of our source code and audio examples are publicly available.

****

Translating Robot Skills: Learning Unsupervised Skill Correspondences Across Robots

Tanmay Shankar,Yixin Lin,Aravind Rajeswaran,Vikash Kumar,Stuart Anderson,Jean Oh

In this paper, we explore how we can endow robots with the ability to learn correspondences between their own skills, and those of morphologically different robots in different domains, in an entirely unsupervised manner. We make the insight that different morphological robots use similar task strategies to solve similar tasks. Based on this insight, we frame learning skill correspondences as a problem of matching distributions of sequences of skills across robots. We then present an unsupervised objective that encourages a learnt skill translation model to match these distributions across domains, inspired by recent advances in unsupervised machine translation. Our approach is able to learn semantically meaningful correspondences between skills across 3 robot domain pairs despite being completely unsupervised. Further, the learnt correspondences enable the transfer of task strategies across robots and domains.

We present dynamic visualizations of our results at https://sites.google.com/view/translatingrobotskills/home.

****

Encoding Hierarchical Information in Neural Networks Helps in Subpopulation Shift

Amitangshu Mukherjee,Isha Garg,Kaushik Roy

Over the past decade, deep neural networks have proven to be adept in image classification tasks, often even surpassing humans in terms of accuracy. However, standard neural networks often fail to understand the concept of hierarchical structures and dependencies among different classes for vision related tasks. Humans on the other hand, seem to learn categories conceptually, progressively growing from understanding high-level concepts down to granular levels of categories. One of the issues arising from the inability of neural networks to encode such dependencies within its learned structure is that of subpopulation shift -- where models are queried with novel unseen classes taken from a shifted population of the training set categories. Since the neural network treats each class as independent from all others, it struggles to categorize shifting populations that are dependent at higher levels of the hierarchy. In this work, we study the aforementioned problems through the lens of a novel conditional supervised training framework. We tackle subpopulation shift by a structured learning procedure that incorporates hierarchical information conditionally through labels. Furthermore, we introduce a notion of graphical distance to model the catastrophic effect of mispredictions. We show that learning in this structured hierarchical manner results in networks that are more robust against subpopulation shifts, with an impro vement of around 2% in terms of accuracy and around 8.5% in terms of graphical distance over standard models on subpopulation shift benchmarks.

****

Text Style Transfer with Confounders

Tianxiao Shen,Regina Barzilay,Tommi S. Jaakkola

Existing methods for style transfer operate either with paired sentences or distributionally matched corpora which differ only in the desired style. In this paper, we relax this restriction and consider data sources with additional confounding differences, from which the desired style needs to be inferred. Specifically, we first learn an invariant style classifier that takes out nuisance variation, and then introduce an orthogonal classifier that highlights the confounding cues. The resulting pair of classifiers guide us to transfer text in the specified direction, creating sentences of the type not seen during training. Experiments show that using positive and negative review datasets from different categories, we can successfully transfer the sentiment without changing the category.

****

Quantized sparse PCA for neural network weight compression

Andrey Kuzmin,Mart Van Baalen,Markus Nagel,Arash Behboodi

In this paper, we introduce a novel method of weight compression. In our method, we store weight tensors as sparse, quantized matrix factors, whose product is computed on the fly during inference to generate the target model's weight tensors. The underlying matrix factorization problem can be considered as a quantized sparse PCA problem and solved through iterative projected gradient descent methods. Seen as a unification of weight SVD, vector quantization and sparse PCA, our method achieves or is on par with state-of-the-art trade-offs between accuracy and model size. Our method is applicable to both moderate compression regime, unlike vector quantization, and extreme compression regime.

****

Mapping conditional distributions for domain adaptation under generalized target shift

Matthieu Kirchmeyer,Alain Rakotomamonjy,Emmanuel de Bezenac,patrick gallinari

We consider the problem of unsupervised domain adaptation (UDA) between a source and a target domain under conditional and label shift a.k.a Generalized Target Shift (GeTarS). Unlike simpler UDA settings, few works have addressed this challenging problem. Recent approaches learn domain-invariant representations, yet they have practical limitations and rely on strong assumptions that may not hold in practice. In this paper, we explore a novel and general approach to align pretrained representations, which circumvents existing drawbacks. Instead of constraining representation invariance, it learns an optimal transport map, implemented as a NN, which maps source representations onto target ones. Our approach is flexible and scalable, it preserves the problem's structure and it has strong theoretical guarantees under mild assumptions. In particular, our solution is unique, matches conditional distributions across domains, recovers target proportions and explicitly controls the target generalization risk. Through an exhaustive comparison on several datasets, we challenge the state-of-the-art in GeTarS.

****

On the Generalization of Models Trained with SGD: Information-Theoretic Bounds and Implications

Ziqiao Wang,Yongyi Mao

This paper follows up on a recent work of Neu et al. (2021) and presents some new information-theoretic upper bounds for the generalization error of machine learning models, such as neural networks, trained with SGD. We apply these bounds to analyzing the generalization behaviour of linear and two-layer ReLU networks. Experimental study of these bounds provide some insights on the SGD training of neural networks. They also point to a new and simple regularization scheme which we show performs comparably to the current state of the art.

****

A theoretically grounded characterization of feature representations

Bharath Hariharan,Cheng Perng Phoo

A large body of work has explored how learned feature representations can be use ful for a variety of downstream tasks. This is true even when the downstream tasks differ greatly from the actual objective used to (pre)train the feature representation. This observation underlies the success of, e.g., few-shot learning, transfer learning and self-supervised learning, among others. However, very little is understood about why such transfer is successful, and more importantly, how one should choose the pre-training task. As a first step towards this understanding, we ask: what makes a feature representation good for a target task? We present simple, intuitive measurements of the feature space that are good predictors of downstream task performance. We present theoretical results showing how these measurements can be used to bound the error of the downstream classifiers, and show empirically that these bounds correlate well with actual downstream performance. Finally, we show that our bounds are practically useful for choosing the right pre-trained representation for a target task.

****

Amortized Implicit Differentiation for Stochastic Bilevel Optimization
Michael Arbel,Julien Mairal

We study a class of algorithms for solving bilevel optimization problems in both stochastic and deterministic settings when the inner-level objective is strongly convex. Specifically, we consider algorithms based on inexact implicit differentiation and we exploit a warm-start strategy to amortize the estimation of the exact gradient. We then introduce a unified theoretical framework inspired by the study of singularly perturbed systems to analyze such amortized algorithms. By using this framework, our analysis shows these algorithms to match the computational complexity of oracle methods that have access to an unbiased estimate of the gradient, thus outperforming many existing results for bilevel optimization. We illustrate these findings on synthetic experiments and demonstrate the efficiency of these algorithms on hyper-parameter optimization experiments involving several thousands of variables.

****

Multi-objective Optimization by Learning Space Partition
Yiyang Zhao,Linnan Wang,Kevin Yang,Tianjun Zhang,Tian Guo,Yuandong Tian

In contrast to single-objective optimization (SOO), multi-objective optimization (MOO) requires an optimizer to find the Pareto frontier, a subset of feasible solutions that are not dominated by other feasible solutions. In this paper, we propose LaMOO, a novel multi-objective optimizer that learns a model from observed samples to partition the search space and then focus on promising regions that are likely to contain a subset of the Pareto frontier. The partitioning is based on the dominance number, which measures "how close'' a data point is to the Pareto frontier among existing samples. To account for possible partition errors due to limited samples and model mismatch, we leverage Monte Carlo Tree Search (MCTS) to exploit promising regions while exploring suboptimal regions that may turn out to contain good solutions later. Theoretically, we prove the efficacy of learning space partitioning via LaMOO under certain assumptions. Empirically, on the HyperVolume (HV) benchmark, a popular MOO metric, LaMOO substantially outperforms strong baselines on multiple real-world MOO tasks, by up to 225% in sample efficiency for neural architecture search on Nasbench201, and up to 10% for molecular design.

****

Mapping Language Models to Grounded Conceptual Spaces
Roma Patel,Ellie Pavlick

A fundamental criticism of text-only language models (LMs) is their lack of grounding---that is, the ability to tie a word for which they have learned a representation, to its actual use in the world. However, despite this limitation, large pre-trained LMs have been shown to have a remarkable grasp of the conceptual structure of language, as demonstrated by their ability to answer questions, generate fluent text, or make inferences about entities, objects, and properties that they have never physically observed. In this work we investigate the extent to which the rich conceptual structure that LMs learn indeed reflects the conceptual structure of the non-linguistic world---which is something that LMs have never observed. We do this by testing whether the LMs can learn to map an entire conc eptual domain (e.g., direction or colour) onto a grounded world representation given only a small number of examples. For example, we show a model what the word ``left" means using a textual depiction of a grid world, and assess how well it can generalise to related concepts, for example, the word ``right", in a similar grid world. We investigate a range of generative language models of varying sizes (including GPT-2 and GPT-3), and see that although the smaller models struggle to perform this mapping, the largest model can not only learn to ground the concepts that it is explicitly taught, but appears to generalise to several instances of unseen concepts as well. Our results suggest an alternative means of building grounded language models: rather than learning grounded representations ``from scratch'', it is possible that large text-only models learn a sufficiently rich conceptual structure that could allow them to be grounded in a data-efficient way.

****

Improving State-of-the-Art in One-Class Classification by Leveraging Unlabeled Data

Farid Bagirov,Dmitry Ivanov,Aleksei Shpilman

Recent advances in One-Class (OC) classification combine the ability to learn exclusively from positive examples with the expressive power of deep neural networks. A cornerstone of OC methods is to make assumptions regarding negative distribution, e.g., that negative data are scattered uniformly or concentrated in the origin. An alternative approach employed in Positive-Unlabeled (PU) learning is to additionally leverage unlabeled data to approximate negative distribution more precisely. In this paper, our goal is to find the best ways to utilize unlabeled data on top of positive data in different settings. While it is reasonable to expect that PU algorithms outperform OC algorithms due to access to more data, we find that the opposite can be true if unlabeled data is unreliable, i.e. contain negative examples that are either too few or sampled from a different distribution. As an alternative to using existing PU algorithms, we propose to modify OC algorithms to incorporate unlabeled data. We find that such PU modifications can consistently benefit even from unreliable unlabeled data if they satisfy a crucial property: when unlabeled data consists exclusively of positive examples, the PU modification becomes equivalent to the original OC algorithm. Our main practical recommendation is to use state-of-the-art PU algorithms when unlabeled data is reliable and to use PU modifications of state-of-the-art OC algorithms that satisfy the formulated property otherwise. Additionally, we make a progress towards distinguishing the cases of reliable and unreliable unlabeled data using statistical tests.

****

The Efficiency Misnomer

Mostafa Dehghani,Yi Tay,Anurag Arnab,Lucas Beyer,Ashish Vaswani

Model efficiency is a critical aspect of developing and deploying machine learning models.

Inference time and latency directly affect the user experience, and some applications have hard requirements. In addition to inference costs, model training also have direct financial and environmental impacts.

Although there are numerous well-established metrics (cost indicators) for measuring model efficiency, researchers and practitioners often assume that these metrics are correlated with each other and report only a few of them.

In this paper, we thoroughly discuss common cost indicators, their advantages and disadvantages, and how they can contradict each other.

We demonstrate how incomplete reporting of cost indicators can lead to partial conclusions and a blurred or incomplete picture of the practical considerations of different models. We further present suggestions to improve reporting of efficiency metrics.

****

Action-Sufficient State Representation Learning for Control with Structural Constraints

Biwei Huang,Chaochao Lu,Liu Leqi,José Miguel Hernández-Lobato,Clark Glymour,Bernhard Schölkopf,Kun Zhang

Perceived signals in real-world scenarios are usually high-dimensional and noisy, and finding and using their representation that contains essential and sufficient information required by downstream decision-making tasks will help improve computational efficiency and generalization ability in the tasks. In this paper, we focus on partially observable environments and propose to learn a minimal set of state representations that capture sufficient information for decision-making, termed \textit{Action-Sufficient state Representations} (ASRs). We build a generative environment model for the structural relationships among variables in the system and present a principled way to characterize ASRs based on structural constraints and the goal of maximizing cumulative reward in policy learning. We then develop a structured sequential Variational Auto-Encoder to estimate the environment model and extract ASRs. Our empirical results on CarRacing and VizDoom demonstrate a clear advantage of learning and using ASRs for policy learning. Moreover, the estimated environment model and ASRs allow learning behaviors from imagined outcomes in the compact latent space to improve sample efficiency.

****

Conditional Generative Quantile Networks via Optimal Transport and Convex Potentials

Jesse Sun,Dihong Jiang,Yaoliang Yu

Quantile regression has a natural extension to generative modelling by leveraging a stronger convergence in pointwise rather than in distribution. While the pinball quantile loss works in the scalar case, it does not have a provable extension to the vector case. In this work, we consider a quantile approach to generative modelling using optimal transport with provable guarantees. We suggest and prove that by optimizing smooth functions with respect to the dual of the correlation maximization problem, the optimum is convex almost surely and hence construct a Brenier map as our generative quantile network. Furthermore, we introduce conditional generative modelling with a Kantorovich dual objective by constructing an affine latent model with respect to the covariates. Through extensive experiments on synthetic and real datasets for conditional generative and probabilistic forecasting tasks, we demonstrate the efficacy and versatility of our theoretically motivated model as a distribution estimator and conditioner.

****

Hybrid Memoised Wake-Sleep: Approximate Inference at the Discrete-Continuous Interface

Tuan Anh Le,Katherine M. Collins,Luke Hewitt,Kevin Ellis,Siddharth N,Samuel Gershman,Joshua B. Tenenbaum

Modeling complex phenomena typically involves the use of both discrete and continuous variables. Such a setting applies across a wide range of problems, from identifying trends in time-series data to performing effective compositional scene understanding in images. Here, we propose Hybrid Memoised Wake-Sleep (HMWS), an algorithm for effective inference in such hybrid discrete-continuous models. Prior approaches to learning suffer as they need to perform repeated expensive inner-loop discrete inference. We build on a recent approach, Memoised Wake-Sleep (MWS), which alleviates part of the problem by memoising discrete variables, and extend it to allow for a principled and effective way to handle continuous variables by learning a separate recognition model used for importance-sampling based approximate inference and marginalization. We evaluate HMWS in the GP-kernel learning and 3D scene understanding domains, and show that it outperforms current state-of-the-art inference methods.

****

Adversarial Retriever-Ranker for Dense Text Retrieval

Hang Zhang,Yeyun Gong,Yelong Shen,Jiancheng Lv,Nan Duan,Weizhu Chen

Current dense text retrieval models face two typical challenges. First, it adopts a siamese dual-encoder architecture to encode query and document independently for fast indexing and searching, whereas neglecting the finer-grained term-wise interactions. This results in a sub-optimal recall performance. Second, it highly relies on a negative sampling technique to build up the negative documents in its contrastive loss. To address these challenges, we present Adversarial Retriever-Ranker (AR2), which consists of a dual-encoder retriever plus a cross-encod er ranker. The two models are jointly optimized according to a minimax adversarial objective: the retriever learns to retrieve negative documents to cheat the ranker, while the ranker learns to rank a collection of candidates including both the ground-truth and the retrieved ones, as well as providing progressive direct feedback to the dual-encoder retriever. Through this adversarial game, the retriever gradually produces harder negative documents to train a better ranker, whereas the cross-encoder ranker provides progressive feedback to improve retriever. We evaluate AR2 on three benchmarks. Experimental results show that AR2 consistently and significantly outperforms existing dense retriever methods and achieves new state-of-the-art results on all of them. This includes the improvements on Natural Questions R@5 to 77.9%(+2.1%), TriviaQA R@5 to 78.2%(+1.4), and MS-MARCO MRR@10 to 39.5%(+1.3%). We will make our code, models, and data publicly available.

****

Pretext Tasks Selection for Multitask Self-Supervised Speech Representation Learning

Salah Zaiem,Titouan Parcollet,Slim Essid,Abdelwahab HEBA

Through solving pretext tasks, self-supervised learning leverages unlabeled data to extract useful latent representations replacing traditional input features in the downstream task. In audio/speech signal processing, a wide range of features where engineered through decades of research efforts. As it turns out, learning to predict such features (a.k.a pseudo-labels) has proven to be a particularly relevant pretext task, leading to useful self-supervised representations which prove to be effective for downstream tasks. However, methods and common practices for combining such pretext tasks for better performance on the downstream task have not been explored and understood properly. In fact, the process relies almost exclusively on a computationally heavy experimental procedure, which becomes intractable with the increase of the number of pretext tasks. This paper introduces a method to select a group of pretext tasks among a set of candidates. The method we propose estimates calibrated weights for the partial losses corresponding to the considered pretext tasks during the self-supervised training process. The experiments conducted on automatic speech recognition, speaker and emotion recognition validate our approach, as the groups selected and weighted with our method perform better than classic baselines, thus facilitating the selection and combination of relevant pseudo-labels for self-supervised representation learning.

****

Fieldwise Factorized Networks for Tabular Data Classification

Chen Almagor,Yedid Hoshen

Tabular data is one of the most common data-types in machine learning, however, deep neural networks have not yet convincingly outperformed classical baselines on such datasets. In this paper, we first investigate the theoretical connection between neural network and factorization machine techniques, and present fieldwise factorized neural networks (F2NN), a neural network architecture framework that is aimed for tabular classification. Our framework learns high-dimensional field representations by a low-rank factorization, and handles both categorical and numerical fields. Furthermore, we show that simply by changing our penultimate activation function, the framework recovers a range of popular tabular classification methods. We evaluate our method against state-of-the-art tabular baselines, including tree-based and deep neural network methods, on a range of tasks. Our findings suggest that our theoretically grounded but simple and shallow neural network architecture achieves as strong or better results than more complex methods.

****

A Closer Look at Smoothness in Domain Adversarial Training

Harsh Rangwani,Sumukh K Aithal,Arihant Jain,Venkatesh Babu Radhakrishnan

Domain adversarial training has been ubiquitous for achieving invariant representations and is used widely for various domain adaptation tasks. In recent times methods converging to smooth optima have shown improved generalization for super vised learning tasks like classification. In this work, we analyze the effect of smoothness enhancing formulations on domain adversarial training, the objective of which is a combination of classification and adversarial terms. In contrast to classification loss, our analysis shows that \textit{converging to smooth minima w.r.t. adversarial loss leads to sub-optimal generalization on the target domain}. Based on the analysis, we introduce the Smooth Domain Adversarial training (SDAT) procedure, which effectively enhances the performance of existing domain adversarial methods for both classification and object detection tasks. Our smoothness analysis also provides insight into the extensive usage of SGD over Adam in domain adversarial training.

****

Conditioning Sequence-to-sequence Networks with Learned Activations

Alberto Gil Couto Pimentel Ramos,Abhinav Mehrotra,Nicholas Donald Lane,Sourav Bhattacharya

Conditional neural networks play an important role in a number of sequence-to-sequence modeling tasks, including personalized sound enhancement (PSE), speaker dependent automatic speech recognition (ASR), and generative modeling such as text-to-speech synthesis. In conditional neural networks, the output of a model is often influenced by a conditioning vector, in addition to the input. Common approaches of conditioning include input concatenation or modulation with the conditioning vector, which comes at a cost of increased model size. In this work, we introduce a novel approach of neural network conditioning by learning intermediate layer activations based on the conditioning vector. We systematically explore and show that learned activation functions can produce conditional models with comparable or better quality, while decreasing model sizes, thus making them ideal candidates for resource-efficient on-device deployment. As exemplary target use-cases we consider (i) the task of PSE as a pre-processing technique for improving telephony or pre-trained ASR performance under noise, and (ii) personalized ASR in single speaker scenarios. We find that conditioning via activation function learning is an effective modeling strategy, suggesting a broad applicability of the proposed technique across a number of application domains.

****

Deep Active Learning with Noise Stability

Xingjian Li,Pengkun Yang,Tianyang Wang,Min Xu,Dejing Dou,Cheng-zhong Xu

Uncertainty estimation for unlabeled data is crucial to active learning. With a deep neural network employed as the backbone model, the data selection process is highly challenged due to the potential over-confidence of the model inference. Existing methods usually resort to multi-pass model training or adversarial training to handle this challenge, resulting in complex and inefficient pipelines, which prevent the deployment in practice. To address such an issue, in this work we propose a novel Single-Training Multi-Inference algorithm that leverages noise stability to estimate data uncertainty. Specifically, it is measured by to what degree the output deviates from the original observation when the model parameters are randomly perturbed by noise. We provide theoretical analysis of using small Gaussian noise, showing that our method has a solid connection with the classical theory of variance reduction, i.e. labelling a data sample of higher uncertainty, indicated by the inverse noise stability, contributes more to reducing the variance of existing training samples. Despite its simplicity and efficiency, our method outperforms the state-of-the-art active learning baselines in image classification and semantic segmentation tasks.

****

LOSSY COMPRESSION WITH DISTRIBUTION SHIFT AS ENTROPY CONSTRAINED OPTIMAL TRANSPORT

Huan Liu,George Zhang,Jun Chen,Ashish J Khisti

We study an extension of lossy compression where the reconstruction distribution is different from the source distribution in order to account for distributional shift due to processing. We formulate this as a generalization of optimal transport with an entropy bottleneck to account for the rate constraint due to compression. We provide expressions for the tradeoff between compression rate and the achievable distortion with and without shared common randomness between the en coder and decoder. We study the examples of binary, uniform and Gaussian sources (in an asymptotic setting) in detail and demonstrate that shared randomness can strictly improve the tradeoff. For the case without common randomness and squared-Euclidean distortion, we show that the optimal solution partially decouples into the problem of optimal compression and transport and also characterize the penalty associated with fully decoupling them. We provide experimental results by training deep learning end-to-end compression systems for performing denoising on SVHN and super-resolution on MNIST suggesting consistency with our theoretical results.

****

A Rate-Distortion Approach to Domain Generalization

Yihang Chen,Grigorios Chrysos,Volkan Cevher

Domain generalization deals with the difference in the distribution between the training and testing datasets, i.e., the domain shift problem, by extracting domain-invariant features. In this paper, we propose an information-theoretic approach for domain generalization. We first establish the domain transformation model, mapping a domain-free latent image into a domain. Then, we cast the domain generalization as a rate-distortion problem, and use the information bottleneck penalty to measure how well the domain-free latent image is reconstructed from a compressed representation of a domain-specific image compared to its direct prediction from the domain-specific image itself. We prove that the information bottleneck penalty guarantees that domain-invariant features can be learned. Lastly, we draw links of our proposed method with self-supervised contrastive learning without negative data pairs. Our empirical study on two different tasks verifies the improvement over recent baselines.

****

Independent SE(3)-Equivariant Models for End-to-End Rigid Protein Docking

Octavian-Eugen Ganea,Xinyuan Huang,Charlotte Bunne,Yatao Bian,Regina Barzilay,Tommi S. Jaakkola,Andreas Krause

Protein complex formation is a central problem in biology, being involved in most of the cell's processes, and essential for applications, e.g. drug design or protein engineering. We tackle rigid body protein-protein docking, i.e., computationally predicting the 3D structure of a protein-protein complex from the individual unbound structures, assuming no conformational change within the proteins happens during binding. We design a novel pairwise-independent SE(3)-equivariant graph matching network to predict the rotation and translation to place one of the proteins at the right docked position relative to the second protein. We mathematically guarantee a basic principle: the predicted complex is always identical regardless of the initial locations and orientations of the two structures. Our model, named EquiDock, approximates the binding pockets and predicts the docking poses using keypoint matching and alignment, achieved through optimal transport and a differentiable Kabsch algorithm. Empirically, we achieve significant running time improvements and often outperform existing docking software despite not relying on heavy candidate sampling, structure refinement, or templates.

****

Transliteration: A Simple Technique For Improving Multilingual Language Modeling

Ibraheem Muhammad Moosa,Mahmud Elahi Akhter,Ashfia Binte Habib

While impressive performance in natural language processing tasks has been achieved for many languages by transfer learning from large pretrained multilingual language models, it is limited by the unavailability of large corpora for most languages and the barrier of different scripts. Script difference forces the tokens of two languages to be separated at the input. Thus we hypothesize that transliterating all the languages to the same script can improve the performance of language models. Languages of South Asia and Southeast Asia present a unique opportunity of testing this hypothesis as almost all of the major languages in this region have their own script. Nevertheless, it is possible to transliterate them to a single representation easily. We validate our hypothesis empirically by pretraining ALBERT models on the Indo-Aryan languages available on the OSCAR corpus and measuring the model's performance on the Indo-Aryan subset of the IndicGLUE benchmark. Compared to the non-transliteration-based model, the transliteration-based model (termed XLM-Indic) shows significant improvement on almost all tasks of IndicGLUE. For example, XLM-Indic performed better on News Classification (0.41%), Multiple Choice QA (4.62%), NER (6.66%), and Cloze-Style QA (3.32%). In addition, XLM-Indic establishes new SOTA results for most tasks the on IndicGLUE benchmark while being competitive at the rest. Across the tasks of IndicGLUE, the most underrepresented languages seem to gain the most improvement. For instance, for the NER, XLM-Indic achieves 10%, 35%, and 58.5% better F1-scores on Gujarati, Panjabi, and Oriya languages compared to the current SOTA.

****

Unsupervised Vision-Language Grammar Induction with Shared Structure Modeling

Bo Wan,Wenjuan Han,Zilong Zheng,Tinne Tuytelaars

We introduce a new task, unsupervised vision-language (VL) grammar induction. Given an image-caption pair, the goal is to extract a shared hierarchical structure for both image and language simultaneously. We argue that such structured output, grounded in both modalities, is a clear step towards the high-level understanding of multimodal information. Besides challenges existing in conventional visually grounded grammar induction tasks, VL grammar induction requires a model to capture contextual semantics and perform a fine-grained alignment. To address these challenges, we propose a novel method, CLIORA, which constructs a shared vision-language constituency tree structure with context-dependent semantics for all possible phrases in different levels of the tree. It computes a matching score between each constituent and image region, trained via contrastive learning. It integrates two levels of fusion, namely at feature-level and at score-level, so as to allow fine-grained alignment. We introduce a new evaluation metric for VL grammar induction, CCRA, and show a 3.3% improvement over a strong baseline on Flickr30k Entities. We also evaluate our model via two derived tasks, i.e., language grammar induction and phrase grounding, and improve over the state-of-the-art for both.

****

Physics Informed Machine Learning of SPH: Machine Learning Lagrangian Turbulence

Michael J Woodward,Yifeng Tian,Criston Hyett,Chris Fryer,Daniel Livescu,Misha Stepanov,Michael Chertkov

Smoothed particle hydrodynamics (SPH) is a mesh-free Lagrangian method for obtaining approximate numerical solutions of the equations of fluid dynamics, which has been widely applied to weakly- and strongly compressible turbulence in astrophysics and engineering applications. We present a learn-able hierarchy of parameterized and "physics-explainable" SPH informed fluid simulators using both physics based parameters and Neural Networks as universal function approximators. Our learning algorithm develops a mixed mode approach, mixing forward and reverse mode automatic differentiation with forward and adjoint based sensitivity analyses to efficiently perform gradient based optimization. We show that our physics informed learning method is capable of: (a) solving inverse problems over the physically interpretable parameter space, as well as over the space of Neural Network parameters; (b) learning Lagrangian statistics of turbulence; (c) combining Lagrangian trajectory based, probabilistic, and Eulerian field based loss functions; and (d) extrapolating beyond training sets into more complex regimes of interest. Furthermore, our hierarchy of models gradually introduces more physical structure, which we show improves interpretability, generalizability (over larger ranges of time scales and Reynolds numbers), preservation of physical symmetries, and requires less training data.

****

Understanding the Generalization Gap in Visual Reinforcement Learning

Anurag Ajay,Ge Yang,Ofir Nachum,Pulkit Agrawal

Deep Reinforcement Learning (RL) agents have achieved superhuman performance on several video game suites. However, unlike humans, the trained policies fail to transfer between related games or even between different levels of the same game. Recent works have attempted to reduce this generalization gap using ideas such as data augmentation and learning domain invariant features. However, the trans fer performance still remains unsatisfactory. In this work, we use procedurally generated video games to empirically investigate several hypotheses to explain the lack of transfer. We also show that simple auxiliary tasks can improve the generalization of policies. Contrary to the belief that policy adaptation to new levels requires full policy finetuning, we find that visual features transfer across levels, and only the parameters, that use these visual features to predict actions, require finetuning. Finally, to inform fruitful avenues for future research, we construct simple oracle methods that close the generalization gap.

****

Equivariant Self-Supervised Learning: Encouraging Equivariance in Representations

Rumen Dangovski,Li Jing,Charlotte Loh,Seungwook Han,Akash Srivastava,Brian Cheung,Pulkit Agrawal,Marin Soljacic

In state-of-the-art self-supervised learning (SSL) pre-training produces semantically good representations by encouraging them to be invariant under meaningful transformations prescribed from human knowledge. In fact, the property of invariance is a trivial instance of a broader class called equivariance, which can be intuitively understood as the property that representations transform according to the way the inputs transform. Here, we show that rather than using only invariance, pre-training that encourages non-trivial equivariance to some transformations, while maintaining invariance to other transformations, can be used to improve the semantic quality of representations. Specifically, we extend popular SSL methods to a more general framework which we name Equivariant Self-Supervised Learning (E-SSL). In E-SSL, a simple additional pre-training objective encourages equivariance by predicting the transformations applied to the input. We demonstrate E-SSL’s effectiveness empirically on several popular computer vision benchmarks, e.g. improving SimCLR to 72.5% linear probe accuracy on ImageNet. Furthermore, we demonstrate usefulness of E-SSL for applications beyond computer vision; in particular, we show its utility on regression problems in photonics science. Our code, datasets and pre-trained models are available at https://github.com/rdangovs/essl to aid further research in E-SSL.

****

Fooling Adversarial Training with Induction Noise

Zhirui Wang,Yifei Wang,Yisen Wang

Adversarial training is widely believed to be a reliable approach to improve model robustness against adversarial attack. However, in this paper, we show that when trained on one type of poisoned data, adversarial training can also be fooled to have catastrophic behavior, e.g., $<1\%$ robust test accuracy with $>90\%$ robust training accuracy on CIFAR-10 dataset. Previously, there are other types of noise poisoned in the training data that have successfully fooled standard training (15.8% standard test accuracy with 99.9% standard training accuracy on CIFAR-10 dataset), but their poisonings can be easily removed when adopting adversarial training. Therefore, we aim to design a new type of inducing noise, named ADVIN, which is an irremovable poisoning of training data. ADVIN can not only degrade the robustness of adversarial training by a large margin, for example, from 51.7% to 0.57% on CIFAR-10 dataset, but also be effective for fooling standard training (13.1% standard test accuracy with 100% standard training accuracy). Additionally, ADVIN can be applied to preventing personal data (like selfies) from being exploited without authorization under whether standard or adversarial training.

****

Gaussian Differential Privacy Transformation: from identification to application

Yi Liu,Ke Sun,Bei Jiang,Linglong Kong

Gaussian differential privacy (GDP) is a single-parameter family of privacy notions that provides coherent guarantees to avoid the exposure of individuals from machine learning models. Relative to traditional (ϵ,δ)-differential privacy (DP), GDP is more interpretable and tightens the bounds given by standard DP composition theorems. In this paper, we start with an exact privacy profile characterization of (ϵ,δ)-DP and then define an efficient, tractable, and visualizable tool, called the Gaussian differential privacy transform ation (GDPT). With theoretical property of the GDPT, we develop an easy-to-verify criterion to characterize and identify GDP algorithms. Based on our criterion, an algorithm is GDP if and only if an asymptotic condition on its privacy profile is met. By development of numerical properties of the GDPT, we give a method to narrow down possible values of an optimal privacy measurement μ with an arbitrarily small and quantifiable margin of error. As applications of our newly developed tools, we revisit some established \ed-DP algorithms and find that their utility can be improved. We additionally make a comparison between two single-parameter families of privacy notions, ϵ-DP and μ-GDP. Lastly, we use the GDPT to examine the effect of subsampling under the GDP framework.

****

Revisiting Out-of-Distribution Detection: A Simple Baseline is Surprisingly Effective

Julian Bitterwolf,Alexander Meinke,Maximilian Augustin,Matthias Hein

It is an important problem in trustworthy machine learning to recognize out-of-distribution (OOD) inputs which are inputs unrelated to the in-distribution task. Many out-of-distribution detection methods have been suggested in recent years. The goal of this paper is to recognize common objectives as well as to identify the implicit scoring functions of different OOD detection methods. In particular, we show that binary discrimination between in- and (different) out-distributions is equivalent to several different formulations of the OOD detection problem. When trained in a shared fashion with a standard classifier, this binary discriminator reaches an OOD detection performance similar to that of Outlier Exposure. Moreover, we show that the confidence loss which is used by Outlier Exposure has an implicit scoring function which differs in a non-trivial fashion from the theoretically optimal scoring function in the case where training and test out-distribution are the same, but is similar to the one used when training with an extra background class. In practice, when trained in exactly the same way, all these methods perform similarly and reach state-of-the-art OOD detection performance.

****

Only tails matter: Average-Case Universality and Robustness in the Convex Regime

Leonardo Cunha,Gauthier Gidel,Fabian Pedregosa,Courtney Paquette,Damien Scieur

Recent works have studied the average convergence properties of first-order optimization methods on distributions of quadratic problems. The average-case framework allows a more fine-grained and representative analysis of convergence than usual worst-case results, in exchange for a more precise hypothesis over the data generating process, namely assuming knowledge of the expected spectral distribution (e.s.d) of the random matrix associated with the problem. In this work, we show that a problem's asymptotic average complexity is determined by the concentration of eigenvalues near the edges of the e.s.d. We argue that having à priori information on this concentration is a more grounded assumption than complete knowledge of the e.s.d., and that basing our analysis on the approximate concentration is effectively a middle ground between the coarseness of the worst-case convergence and this more unrealistic hypothesis. We introduce the Generalized Chebyshev method, asymptotically optimal under a hypothesis on this concentration, and globally optimal when the e.s.d. follows a Beta distribution. We compare its performance to classical optimization algorithms, such as Gradient Descent or Nesterov's scheme, and we show that, asymptotically, Nesterov's method is universally nearly-optimal in the average-case.

****

Provable Learning of Convolutional Neural Networks with Data Driven Features

Alon Brutzkus,Amir Globerson,eran malach,Shai Shalev-Shwartz

Convolutional networks (CNN) are computationally hard to learn. In practice, however, CNNs are learned successfully on natural image data. In this work, we study a semi-supervised algorithm, that learns a linear classifier over data-dependent features which were obtained from unlabeled data. We show that the algorithm provably learns CNNs, under some natural distributional assumptions. Specifically, it efficiently learns CNNs, assuming the distribution of patches in the input images has low-dimensional structure (e.g., when the patches are sampled from a low-dimensional manifold). We complement our result with a lower bound, showing that the dependence of our algorithm on the dimension of the patch distribution is essentially optimal.

****

Genome Sequence Reconstruction Using Gated Graph Convolutional Network

Lovro Vr■ek,Robert Vaser,Thomas Laurent,Mile Sikic,Xavier Bresson

A quest to determine the human DNA sequence from telomere to telomere started three decades ago and was finally finished in 2021. This accomplishment was a result of a tremendous effort of numerous experts with an abundance of data, various tools, and often included manual inspection during genome reconstruction. Therefore, such method could hardly be used as a general approach to assembling genomes, especially when the assembly speed is important. Motivated by this achievement and aspiring to make it more accessible, we investigate a previously untaken path of applying geometric deep learning to the central part of the genome assembly---untangling a large assembly graph from which a genomic sequence needs to be reconstructed. A graph convolutional network is trained on a dataset generated from human genomic data to reconstruct the genome by finding a path through the assembly graph. We show that our model can compute scores from the lengths of the overlaps between the sequences and the graph topology which, when traversed with a greedy search algorithm, outperforms the greedy search over the overlap lengths only. Moreover, our method reconstructs the correct path through the graph in the fraction of time required for the state-of-the-art de novo assemblers. This favourable result paves the way for the development of powerful graph machine learning algorithms that can solve the de novo genome assembly problem much quicker and possibly more accurately than human handcrafted techniques.

****

Subpixel object segmentation using wavelets and multiresolution analysis

Ray Sheombarsing,Nikita Moriakov,Jan-jakob Sonke,Jonas Teuwen

We propose a novel deep learning framework for fast prediction of boundaries of two-dimensional simply connected domains using wavelets and Multiresolution Analysis (MRA). The boundaries are modelled as (piecewise) smooth closed curves using wavelets and the so-called Pyramid Algorithm. Our network architecture is a hybrid analog of the U-Net, where the down-sampling path is a two-dimensional encoder with learnable filters, and the upsampling path is a one-dimensional decoder, which builds curves up from low to high resolution levels. Any wavelet basis induced by a MRA can be used. This flexibility allows for incorporation of priors on the smoothness of curves. The effectiveness of the proposed method is demonstrated by delineating boundaries of simply connected domains (organs) in medical images using Debauches wavelets and comparing performance with a U-Net baseline. Our model demonstrates up to 5x faster inference speed compared to the U-Net, while maintaining similar performance in terms of Dice score and Hausdorff distance.

****

Pseudo Knowledge Distillation: Towards Learning Optimal Instance-specific Label Smoothing Regularization

Peng Lu,Ahmad Rashid,Ivan Kobyzev,Mehdi Rezagholizadeh,Philippe Langlais

Knowledge Distillation (KD) is an algorithm that transfers the knowledge of a trained, typically larger, neural network into another model under training. Although a complete understanding of KD is elusive, a growing body of work has shown that the success of both KD and label smoothing comes from a similar regularization effect of soft targets. In this work, we propose an instance-specific label smoothing technique, Pseudo-KD, which is efficiently learnt from the data. We devise a two-stage optimization problem that leads to a deterministic and interpretable solution for the optimal label smoothing. We show that Pseudo-KD can be equivalent to an efficient variant of self-distillation techniques, without the need to store the parameters or the output of a trained model. Finally, we conduct experiments on multiple image classification (CIFAR-10 and CIFAR-100) and natural language understanding datasets (the GLUE benchmark) across various neural network architectures and demonstrate that our method is competitive against strong baselines.

****

Direct then Diffuse: Incremental Unsupervised Skill Discovery for State Covering and Goal Reaching

Pierre-Alexandre Kamienny,Jean Tarbouriech,sylvain lamprier,Alessandro Lazaric,Ludovic Denoyer

Learning meaningful behaviors in the absence of reward is a difficult problem in reinforcement learning. A desirable and challenging unsupervised objective is to learn a set of diverse skills that provide a thorough coverage of the state space while being directed, i.e., reliably reaching distinct regions of the environment. In this paper, we build on the mutual information framework for skill discovery and introduce UPSIDE, which addresses the coverage-directedness trade-off in the following ways: 1) We design policies with a decoupled structure of a directed skill, trained to reach a specific region, followed by a diffusing part that induces a local coverage. 2) We optimize policies by maximizing their number under the constraint that each of them reaches distinct regions of the environment (i.e., they are sufficiently discriminable) and prove that this serves as a lower bound to the original mutual information objective. 3) Finally, we compose the learned directed skills into a growing tree that adaptively covers the environment. We illustrate in several navigation and control environments how the skills learned by UPSIDE solve sparse-reward downstream tasks better than existing baselines.

****

Towards a Unified View of Parameter-Efficient Transfer Learning

Junxian He,Chunting Zhou,Xuezhe Ma,Taylor Berg-Kirkpatrick,Graham Neubig

Fine-tuning large pretrained language models on downstream tasks has become the de-facto learning paradigm in NLP. However, conventional approaches fine-tune all the parameters of the pretrained model, which becomes prohibitive as the model size and the number of tasks grow. Recent work has proposed a variety of parameter-efficient transfer learning methods that only fine-tune a small number of (extra) parameters to attain strong performance. While effective, the critical ingredients for success and the connections among the various methods are poorly understood. In this paper, we break down the design of state-of-the-art parameter-efficient transfer learning methods and present a unified framework that establishes connections between them. Specifically, we re-frame them as modifications to specific hidden states in pretrained models, and define a set of design dimensions along which different methods vary, such as the function to compute the modification and the position to apply the modification. Through comprehensive empirical studies across machine translation, text summarization, language understanding, and text classification benchmarks, we utilize the unified view to identify important design choices in previous methods. Furthermore, our unified framework enables the transfer of design elements across different approaches, and as a result we are able to instantiate new parameter-efficient fine-tuning methods that tune less parameters than previous methods while being more effective, achieving comparable results to fine-tuning all parameters on all four tasks.

****

To Impute or Not To Impute? Missing Data in Treatment Effect Estimation

Jeroen Berrevoets,Fergus Imrie,Trent Kyono,James Jordon,Mihaela van der Schaar

Missing data is a systemic problem in practical scenarios that causes noise and bias when estimating treatment effects. This makes treatment effect estimation from data with missingness a particularly tricky endeavour. A key reason for this is that standard assumptions on missingness are rendered insufficient due to the presence of an additional variable, treatment, besides the individual and the outcome. Having a treatment variable introduces additional complexity with respect to why some variables are missing that is overlooked by previous work. In our work we identify a new missingness mechanism, which we term mixed confounded missingness (MCM), where some missingness determines treatment selection and other missingness is determined by treatment selection. Given MCM, we show that naively imputing all data leads to poor performing treatment effects models, as the act of imputation effectively removes information necessary to provide unbiased estimates. However, no imputation at all also leads to biased estimates, as miss ingness determined by treatment divides the population in distinct subpopulations, where estimates across these populations will be biased. Our solution is selective imputation, where we use insights from MCM to inform precisely which variables should be imputed and which should not. We empirically demonstrate how various learners benefit from selective imputation compared to other solutions for missing data.

****

Instance-Adaptive Video Compression: Improving Neural Codecs by Training on the Test Set

Ties van Rozendaal,Johann Brehmer,Yunfan Zhang,Reza Pourreza,Taco Cohen

We introduce a video compression algorithm based on instance-adaptive learning. On each video sequence to be transmitted, we finetune a pretrained compression model. The optimal parameters are transmitted to the receiver along with the latent code. By entropy-coding the parameter updates under a suitable mixture model prior, we ensure that the network parameters can be encoded efficiently. This instance-adaptive compression algorithm is agnostic about the choice of base model and has the potential to improve any neural video codec. On UVG, HEVC, and Xiph datasets, our codec improves the performance of a low-latency scale-space flow model by between 24% and 26% BD-rate savings, and that of a state-of-the-art B-frame model by 17 to 20% BD-rate savings. We also demonstrate that instance-adaptive finetuning improves the robustness to domain shift. Finally, our approach reduces the capacity requirements on compression models. We show that it enables a state-of-the-art performance even after reducing the network size by 72%.

****

Normalization of Language Embeddings for Cross-Lingual Alignment

Prince Osei Aboagye,Yan Zheng,Chin-Chia Michael Yeh,Junpeng Wang,Wei Zhang,Liang Wang,Hao Yang,Jeff Phillips

Learning a good transfer function to map the word vectors from two languages into a shared cross-lingual word vector space plays a crucial role in cross-lingual NLP. It is useful in translation tasks and important in allowing complex models built on a high-resource language like English to be directly applied on an aligned low resource language. While Procrustes and other techniques can align language models with some success, it has recently been identified that structural differences (for instance, due to differing word frequency) create different profiles for various monolingual embedding. When these profiles differ across languages, it correlates with how well languages can align and their performance on cross-lingual downstream tasks. In this work, we develop a very general language embedding normalization procedure, building and subsuming various previous approaches, which removes these structural profiles across languages without destroying their intrinsic meaning. We demonstrate that meaning is retained and alignment is improved on similarity, translation, and cross-language classification tasks. Our proposed normalization clearly outperforms all prior approaches like centering and vector normalization on each task and with each alignment approach.

****

Invariant Learning with Partial Group Labels

Vishnu Suresh Lokhande,Kihyuk Sohn,Jinsung Yoon,Madeleine Udell,Chen-Yu Lee,Tomas Pfister

Learning invariant representations is an important requirement in training machine learning models that are driven by spurious correlations in the datasets. These spurious correlations, between input samples and the target labels, wrongly direct the neural network predictions resulting in poor performance on certain groups, especially the minority groups. Robust training against these spurious correlations requires the knowledge of group membership for every sample. Such a requirement is impractical in situations where the data labelling efforts for minority or rare groups is significantly laborious or where the individuals comprising the dataset choose to conceal sensitive information. On the other hand, the presence of such data collection efforts result in datasets that contain partially labelled group information. Recent works have tackled the fully unsupervised scenario where no labels for groups are available. Thus, we aim to fill the missi ng gap in the literature by tackling a more realistic setting that can leverage partially available sensitive or group information during training. First, we construct a constraint set and derive a high probability bound for the group assignment to belong to the set. Second, we propose an algorithm that optimizes for the worst-off group assignments from the constraint set. Through experiments on image and tabular datasets, we show improvements in the minority group’s performance while preserving overall aggregate accuracy across groups.

****

A neural network framework for learning Green's function

Guochang Lin,Fukai Chen,Pipi Hu,Xiang Chen,Junqing Chen,Jun Wang,Zuoqiang Shi

Green's function plays a significant role in both theoretical analysis and numerical computing of partial differential equations (PDEs). However, in most cases, Green's function is difficult to compute. The troubles arise in the following three folds. Firstly, compared with the original PDE, the dimension of Green's function is doubled, making it impossible to be handled by traditional mesh-based methods. Secondly, Green's function usually contains singularities which increase the difficulty to get a good approximation. Lastly, the computational domain may be very complex or even unbounded. To override these problems, we leverage the fundamental solution, boundary integral method and neural networks to develop a new method for computing Green's function with high accuracy in this paper. We focus on Green's function of Poisson and Helmholtz equations in bounded domains, unbounded domains and domains with interfaces. Extensive experiments illustrate the efficiency and the accuracy of our method for solving Green's function. In addition, we also use the Green's function calculated by our method to solve a class of PDE, and also obtain high-precision solutions, which shows the good generalization ability of our method on solving PDEs.

****

Boosting the Certified Robustness of L-infinity Distance Nets

Bohang Zhang,Du Jiang,Di He,Liwei Wang

Recently, Zhang et al. (2021) developed a new neural network architecture based on ℓ_∞-distance functions, which naturally possesses certified ℓ_∞ robustness by its construction. Despite the novel design and theoretical foundation, so far the model only achieved comparable performance to conventional networks. In this paper, we make the following two contributions: $\mathrm{(i)}$ We demonstrate that ℓ_∞-distance nets enjoy a fundamental advantage in certified robustness over conventional networks (under typical certification approaches); $\mathrm{(ii)}$ With an improved training process we are able to significantly boost the certified accuracy of ℓ_∞-distance nets. Our training approach largely alleviates the optimization problem that arose in the previous training scheme, in particular, the unexpected large Lipschitz constant due to the use of a crucial trick called \textit{ℓ_p-relaxation}. The core of our training approach is a novel objective function that combines scaled cross-entropy loss and clipped hinge loss with a decaying mixing coefficient. Experiments show that using the proposed training strategy, the certified accuracy of ℓ_∞-distance net can be dramatically improved from 33.30% to 40.06% on CIFAR-10 ($\epsilon=8/255$), meanwhile outperforming other approaches in this area by a large margin. Our results clearly demonstrate the effectiveness and potential of ℓ_∞-distance net for certified robustness. Codes are available at https://github.com/zbh2047/L_inf-dist-net-v2.

****

LRN: Limitless Routing Networks for Effective Multi-task Learning

Ryan Wickman,Xiaofei Zhang,Weizi Li

Multi-task learning (MTL) is a field involved with learning multiple tasks simultaneously typically through using shared model parameters. The shared representation enables generalized parameters that are task invariant and assists in learning tasks with sparse data. However, the presence of unforeseen task interference can cause one task to improve at the detriment of another. A recent paradigm constructed to tackle these types of problems is the routing network, that builds neural network architectures from a set of modules conditioned on the input instance, task, and previous output of other modules. This approach has many constr aints, so we propose the Limitless Routing Network (LRN) which removes the constraints through the usage of a transformer-based router and a reevaluation of the state and action space. We also provide a simple solution to the module collapse problem and display superior accuracy performance over several MTL benchmarks compared to the original routing network.

****

Adaptive Cross-Layer Attention for Image Restoration

Yancheng Wang,Yingzhen Yang,Chong Chen,Ning Xu

Non-local attention module has been proven to be crucial for image restoration. Conventional non-local attention processes features of each layer separately, so it risks missing correlation between features among different layers. To address this problem, we propose Cross-Layer Attention (CLA) module in this paper. Instead of finding correlated key pixels within the same layer, each query pixel is allowed to attend to key pixels at previous layers of the network. In order to mitigate the expensive computational cost of such hierarchical attention design, only a small fixed number of keys can be selected for each query from a previous layer. We further propose a variant of CLA termed Adaptive Cross-Layer Attention (ACLA). In ACLA, the number of keys to be aggregated for each query is dynamically selected. A neural architecture search method is used to find the insert positions of ACLA modules to render a compact neural network with compelling performance. Extensive experiments on image restoration tasks including single image super-resolution, image denoising, image demosaicing, and image compression artifacts reduction validate the effectiveness and efficiency of ACLA.

****

Motion Planning Transformers: One Model to Plan them All

Jacob John Johnson,Linjun Li,Ahmed Qureshi,Michael C. Yip

Transformers have become the powerhouse of natural language processing and recently found use in computer vision tasks. Their effective use of attention can be used in other contexts as well, and in this paper, we propose a transformer-based approach for efficiently solving complex motion planning problems. Traditional neural network-based motion planning uses convolutional networks to encode the planning space, but these methods are limited to fixed map sizes, which is often not realistic in the real world. Our approach first identifies regions on the map using transformers to provide attention to map areas likely to include the best path and then applies traditional planners to generate the final collision-free path. We validate our method on a variety of randomly generated environments with different map sizes, demonstrating reduction in planning complexity and achieving comparable accuracy to traditional planners.

****

SparRL: Graph Sparsification via Deep Reinforcement Learning

Ryan Wickman,Xiaofei Zhang,Weizi Li

Graph sparsification concerns data reduction where an edge-reduced graph of a similar structure is preferred. Existing methods are mostly sampling-based, which introduce high computation complexity in general and lack of flexibility for a different reduction objective. We present SparRL, the first general and effective reinforcement learning-based framework for graph sparsification. SparRL can easily adapt to different reduction goals and promise graph-size-independent complexity. Extensive experiments show that SparRL outperforms all prevailing sparsification methods in producing high-quality sparsified graphs concerning a variety of objectives. As graph representations are very versatile, SparRL carries the potential for a broad impact.

****

Rewardless Open-Ended Learning (ROEL)

Alexander Quessy,Thomas Stuart Richardson

Open-ended learning algorithms aim to automatically generate challenges and solutions to an unending sequence of learning opportunities. In Reinforcement Learning (RL) recent approaches to open-ended learning, such as Paired Open-Ended Trailblazer (POET), focus on collecting a diverse set of solutions based on the novelty of an agents pre-defined reward function. In many practical RL tasks definin g an effective reward function a priori is often hard and can hinder an agents ability to explore many behaviors that could ultimately be more performant. In this work we combine open-ended learning with unsupervised reinforcement learning to train agents to learn a diverse set of complex skills. We propose a procedure to combine skill-discovery via mutual information, using the POET algorithm as an open-ended framework to teach agents increasingly complex groups of diverse skills. Experimentally we demonstrate this approach yields agents capable of demonstrating identifiable skills over a range of environments, that can be extracted and utilized to solve a variety of tasks.

****

The Needle in the haystack: Out-distribution aware Self-training in an Open-World Setting

Maximilian Augustin,Matthias Hein

Traditional semi-supervised learning (SSL) has focused on the closed world assumption where all unlabeled samples are task-related. In practice, this assumption is often violated when leveraging data from very large image databases that contain mostly non-task-relevant samples. While standard self-training and other established methods fail in this open-world setting, we demonstrate that our out-distribution-aware self-learning (ODST) with a careful sample selection strategy can leverage unlabeled datasets with millions of samples, more than 1600 times larger than the labeled datasets, and which contain only about 2% task-relevant inputs. Standard and open world SSL techniques degrade in performance when the ratio of task-relevant sample decreases and show a significant distribution shift which is problematic regarding AI safety while ODST outperforms them with respect to test performance, corruption robustness and out-of-distribution detection.

****

Multi-Agent Reinforcement Learning with Shared Resource in Inventory Management

Mingxiao Feng,Guozi Liu,Li Zhao,Lei Song,Jiang Bian,Tao Qin,Wengang Zhou,Houqiang Li,Tie-Yan Liu

We consider inventory management (IM) problem for a single store with a large number of SKUs (stock keeping units) in this paper, where we need to make replenishment decisions for each SKU to balance its supply and demand. Each SKU should cooperate with each other to maximize profits, as well as compete for shared resources e.g., warehouse spaces, budget etc. Co-existence of cooperation and competition behaviors makes IM a complicate game, hence IM can be naturally modelled as a multi-agent reinforcement learning (MARL) problem. In IM problem, we find that agents only interact indirectly with each other through some shared resources, e.g., warehouse spaces. To formally model MARL problems with above structure, we propose shared resource stochastic game along with an efficient algorithm to learn policies particularly for a large number of agents. By leveraging shared-resource structure, our method can greatly reduce model complexity and accelerate learning procedure compared with standard MARL algorithms, as shown by extensive experiments.

****

Representing value functions in power systems using parametric network series

Ruben Chaer,Ximena Caporale,Vanina Camacho,Ignacio Ramírez

We describe a novel architecture for modeling the cost-to-go function in approximate dynamic programming problems involving country-scale, real-life electrical power generation systems. Our particular scenario features a heterogeneous power grid including dozens of renewable energy plants as well as traditional ones; the corresponding state space is in the order of thousands of variables of different types and ranges. While Artificial Neural Networks are a natural choice for modeling such complex cost functions, their effective use hinges on exploiting the particular structure of the problem which, in this case, involves seasonal patterns at many different levels (day, week, year). Our proposed model consists of a series of neural networks whose parameters are themselves parametric functions of a time variable. The parameters of such functions are learned during training along with the network parameters themselves. The new method is shown to out perform the standard backward dynamic programming program currently in use, both in terms of the objective function (total cost of operation over a period) and computational cost. Last, but not least, the resulting model is readily interpretable in terms of the parameters of the learned functions, which capture general trends of the problem, providing useful insight for future improvements.

****

Task-oriented Dialogue System for Automatic Disease Diagnosis via Hierarchical Reinforcement Learning

Kangenbei Liao,CHENG ZHONG,Wei Chen,Qianlong Liu,zhongyu wei,Baolin Peng,Xuanjing Huang

In this paper, we focus on automatic disease diagnosis with reinforcement learning (RL) methods in task-oriented dialogues setting. Different from conventional RL tasks, the action space for disease diagnosis (i.e., symptoms) is inevitably large, especially when the number of diseases increases. However, existing approaches to this problem typically works well in simple tasks but has significant challenges in complex scenarios. Inspired by the offline consultation process, we propose to integrate a hierarchical policy of two levels into the dialogue policy learning. The high level policy consists of a master model that is responsible for triggering a low level model, the low level policy consists of several symptom checkers and a disease classifier. Experimental results on both self-constructed real-world and synthetic datasets demonstrate that our hierarchical framework achieves higher accuracy and symptom recall in disease diagnosis compared with existing systems.

****

GAETS: A Graph Autoencoder Time Series Approach Towards Battery Parameter Estimation

Edward Elson Kosasih,Rucha Bhalchandra Joshi,Janamejaya Channegowda

Lithium-ion batteries are powering the ongoing transportation electrification revolution. Lithium-ion batteries possess higher energy density and favourable electrochemical properties which make it a preferable energy source for electric vehicles. Precise estimation of battery parameters (Charge capacity, voltage etc) is vital to estimate the available range in an electric vehicle. Graph-based estimation techniques enable us to understand the variable dependencies underpinning them to improve estimates. In this paper we employ Graph Neural Networks for battery parameter estimation, we introduce a unique graph autoencoder time series estimation approach. Variables in battery measurements are known to have an underlying relationship with each other in a certain causal structure. Therefore, we include ideas from the field of causal structure learning as a regularisation to our learned adjacency matrix technique. We use graph autoencoder based on a non-linear version of NOTEARS Zheng et al. (2018) as this allowed us to perform gradient-descent in learning the structure (instead of treating it as a combinatorial optimisation problem). The proposed architecture outperforms the state-of-the-art Graph Time Series (GTS) Shang et al. (2021a)   architecture for battery parameter estimation. We call our method GAETS (Graph AutoEncoder Time Series).

****

FitVid: High-Capacity Pixel-Level Video Prediction

Mohammad Babaeizadeh,Mohammad Taghi Saffar,Suraj Nair,Sergey Levine,Chelsea Finn,Dumitru Erhan

An agent that is capable of predicting what happens next can perform a variety of tasks through planning with no additional training. Furthermore, such an agent can internally represent the complex dynamics of the real-world and therefore can acquire a representation useful for a variety of visual perception tasks. This makes predicting the future frames of a video, conditioned on the observed past and potentially future actions, an interesting task which remains exceptionally challenging despite many recent advances. Existing video prediction models have shown promising results on simple narrow benchmarks but they generate low quality predictions on real-life datasets with more complicated dynamics or broader domain. There is a growing body of evidence that underfitting on the training data is one of the primary causes for the low quality predictions. In this paper, we argue that the inefficient use of parameters in the current video models is the main reason for underfitting. Therefore, we introduce a new architecture, named FitVid, which is capable of fitting the common benchmarks so well that it begins to suffer from overfitting -- while having similar parameter count as the current state-of-the-art models. We analyze the consequences of overfitting, illustrating how it can produce unexpected outcomes such as generating high quality output by repeating the training data, and how it can be mitigated using existing image augmentation techniques. As a result, FitVid outperforms the current state-of-the-art models across four different video prediction benchmarks on four different metrics.

****

Generative Posterior Networks for Approximately Bayesian Epistemic Uncertainty Estimation

Melrose Roderick,Felix Berkenkamp,Fatemeh Sheikholeslami,J Zico Kolter

Ensembles of neural networks are often used to estimate epistemic uncertainty in high-dimensional problems because of their scalability and ease of use. These methods, however, are expensive to sample from as each sample requires a new neural network to be trained from scratch. We propose a new method, Generative Posterior Networks (GPNs), a generative model that, given a prior distribution over functions, approximates the posterior distribution directly by regularizing the network towards samples from the prior. This allows our method to quickly sample from the posterior and construct confidence bounds. We prove theoretically that our method indeed approximates the Bayesian posterior and show empirically that it improves epistemic uncertainty estimation over competing methods.

****

FlowX: Towards Explainable Graph Neural Networks via Message Flows

Shurui Gui,Hao Yuan,Jie Wang,Qicheng Lao,Kang Li,Shuiwang Ji

We investigate the explainability of graph neural networks (GNNs) as a step towards elucidating their working mechanisms. While most current methods focus on explaining graph nodes, edges, or features, we argue that, as the inherent functional mechanism of GNNs, message flows are more natural for performing explainability. To this end, we propose a novel method here, known as FlowX, to explain GNNs by identifying important message flows. To quantify the importance of flows, we propose to employ the concept of Shapley values from cooperative game theory. To tackle the complexity of computing Shapley values, we propose an approximation scheme to compute Shapley values as initial assessments of flow importance. We then propose a learning algorithm to refine scores and improve explainability. Experimental studies on both synthetic and real-world datasets demonstrate that our proposed FlowX leads to improved explainability of GNNs.

****

On Learning with Fairness Trade-Offs

Francois Buet-Golfouse

Previous literature has shown that bias mitigating algorithms were sometimes prone to overfitting and had poor out-of-sample generalisation. This paper is first and foremost concerned with establishing a mathematical framework to tackle the specific issue of generalisation. Throughout this work, we consider fairness trade-offs and objectives mixing statistical loss over the whole sample and fairness penalties on categories (which could stem from different values of protected attributes), encompassing partial de-biasing. We do so by adopting two different but complementary viewpoints: first, we consider a PAC-type setup and derive probabilistic upper bounds involving sample-only information; second, we leverage an asymptotic framework to derive a closed-form limiting distribution for the difference between the empirical trade-off and the true trade-off. While these results provide guarantees for learning fairness metrics across categories, they also point out to the key (but asymmetric) role played by class imbalance. To summarise, learning fairness without having access to enough category-level samples is hard, and a simple numerical experiment shows that it can lead to spurious results.

****

An Empirical Investigation of the Role of Pre-training in Lifelong Learning

Sanket Vaibhav Mehta,Darshan Patil,Sarath Chandar,Emma Strubell

The lifelong learning paradigm in machine learning is an attractive alternative to the more prominent isolated learning scheme not only due to its resemblance to biological learning, but also its potential to reduce energy waste by obviating excessive model re-training. A key challenge to this paradigm is the phenomenon of catastrophic forgetting. With the increasing popularity and success of pre-trained models in machine learning, we pose the question: What role does pre-training play in lifelong learning, specifically with respect to catastrophic forgetting? We investigate existing methods in the context of large, pre-trained models and evaluate their performance on a variety of text and image classification tasks, including a large-scale study using a novel dataset of 15 diverse NLP tasks. Across all settings, we observe that generic pre-training implicitly alleviates the effects of catastrophic forgetting when learning multiple tasks sequentially compared to randomly initialized models. We then further investigate why pre-training alleviates forgetting in this setting. We study this phenomenon by analyzing the loss landscape, finding that pre-trained weights appear to ease forgetting by leading to wider minima. Based on this insight, we propose jointly optimizing for current task loss and loss basin sharpness in order to explicitly encourage wider basins during sequential fine-tuning. We show that this optimization approach leads to performance comparable to the state-of-the-art in task-sequential continual learning across multiple settings, without retaining a memory that scales in size with the number of tasks.

****

Stochastic Training is Not Necessary for Generalization

Jonas Geiping,Micah Goldblum,Phil Pope,Michael Moeller,Tom Goldstein

It is widely believed that the implicit regularization of SGD is fundamental to the impressive generalization behavior we observe in neural networks. In this work, we demonstrate that non-stochastic full-batch training can achieve comparably strong performance to SGD on CIFAR-10 using modern architectures. To this end, we show that the implicit regularization of SGD can be completely replaced with explicit regularization. Our observations indicate that the perceived difficulty of full-batch training may be the result of its optimization properties and the disproportionate time and effort spent by the ML community tuning optimizers and hyperparameters for small-batch training.

****

Transfer RL across Observation Feature Spaces via Model-Based Regularization

Yanchao Sun,Ruijie Zheng,Xiyao Wang,Andrew E Cohen,Furong Huang

In many reinforcement learning (RL) applications, the observation space is specified by human developers and restricted by physical realizations, and may thus be subject to dramatic changes over time (e.g. increased number of observable features). However, when the observation space changes, the previous policy will likely fail due to the mismatch of input features, and another policy must be trained from scratch, which is inefficient in terms of computation and sample complexity. Following theoretical insights, we propose a novel algorithm which extracts the latent-space dynamics in the source task, and transfers the dynamics model to the target task to use as a model-based regularizer. Our algorithm works for drastic changes of observation space (e.g. from vector-based observation to image-based observation), without any inter-task mapping or any prior knowledge of the target task. Empirical results show that our algorithm significantly improves the efficiency and stability of learning in the target task.

****

Continual Learning of Neural Networks for Realtime Wireline Cable Position Inference

Jun Wang,Tianxiang Su

In the oil fields, Wireline cable is spooled onto a drum where computer vision techniques based on convolutional neural networks (CNNs) are applied to estimate the cable position in real time for automated spooling control. However, as new training data keeps arriving to continuously improve the network, the re-training procedure faces challenges. Online learning fashion with no memory to historical data leads to catastrophic forgetting. Meanwhile, saving all data will cause the disk space and training time to increase without bounds. In this paper, we proposed a method called the modified-REMIND (mREMIND) network. It is a replay-based continual learning method with longer memory to historical data and no memory overflow issues. Information of old data are kept for multiple iterations using a new dictionary update rule. Additionally, by dynamically partitioning the dataset, the method can be applied on devices with limited memory. In our experiments, we compared the proposed method with multiple state-of-the-art continual learning methods and the mREMIND network outperformed others both in accuracy and in disk space usage.

****

Benchmarking Machine Learning Robustness in Covid-19 Spike Sequence Classification

Sarwan Ali,Bikram Sahoo,Pin-Yu Chen,Murray Patterson

The rapid spread of the COVID-19 pandemic has resulted in an unprecedented amount of sequence data of the SARS-CoV-2 viral genome --- millions of sequences and counting. This amount of data, while being orders of magnitude beyond the capacity of traditional approaches to understanding the diversity, dynamics and evolution of viruses, is nonetheless a rich resource for machine learning (ML) and deep learning (DL) approaches as alternatives for extracting such important information from these data. It is of hence utmost importance to design a framework for testing and benchmarking the robustness of these ML and DL approaches.

This paper the first (to our knowledge) to explore such a framework. In this paper, we introduce several ways to perturb SARS-CoV-2 spike protein sequences in ways that mimic the error profiles of common sequencing platforms such as Illumina and PacBio. We show from experiments on a wide array of ML approaches from naive Bayes to logistic regression, that DL approaches are more robust (and accurate) to such adverarial attacks to the input sequences, while k-mer based feature vector representations are more robust than the baseline one-hot embedding.
Our benchmarking framework may developers of futher ML and DL techniques to properly assess their approaches towards understanding the behaviour of the SARS-CoV-2 virus, or towards avoiding possible future pandemics.

****

DIGRAC: Digraph Clustering Based on Flow Imbalance

Yixuan He,Gesine Reinert,Mihai Cucuringu

Node clustering is a powerful tool in the analysis of networks. We introduce a graph neural network framework to obtain node embeddings for directed networks in a self-supervised manner, including a novel probabilistic imbalance loss, which can be used for network clustering. Here, we propose directed flow imbalance measures, which are tightly related to directionality, to reveal clusters in the network even when there is no density difference between clusters. In contrast to standard approaches in the literature, in this paper, directionality is not treated as a nuisance, but rather contains the main signal. DIGRAC optimizes directed flow imbalance for clustering without requiring label supervision, unlike existing GNN methods, and can naturally incorporate node features, unlike existing spectral methods. Experimental results on synthetic data, in the form of directed stochastic block models, and real-world data at different scales, demonstrate that our method, based on flow imbalance, attains state-of-the-art results on directed graph clustering, for a wide range of noise and sparsity levels and graph structures and topologies.

****

GATSBI: Generative Adversarial Training for Simulation-Based Inference

Poornima Ramesh,Jan-Matthis Lueckmann,Jan Boelts,Álvaro Tejero-Cantero,David S. Greenberg,Pedro J. Goncalves,Jakob H. Macke

Simulation-based inference (SBI) refers to statistical inference on stochastic models for which we can generate samples, but not compute likelihoods.
Like SBI algorithms, generative adversarial networks (GANs) do not require explicit likelihoods. We study the relationship between SBI and GANs, and introduce GATSBI, an adversarial approach to SBI. GATSBI reformulates the variational objective in an adversarial setting to learn implicit posterior distributions. Infere

nce with GATSBI is amortised across observations, works in high-dimensional posterior spaces and supports implicit priors. We evaluate GATSBI on two common SBI benchmark problems and on two high-dimensional simulators. On a model for wave propagation on the surface of a shallow water body, we show that GATSBI can return well-calibrated posterior estimates even in high dimensions.

On a model of camera optics, it infers a high-dimensional posterior given an implicit prior, and performs better than a state-of-the-art SBI approach. We also show how GATSBI can be extended to perform sequential posterior estimation to focus on individual observations.

Overall, GATSBI opens up opportunities for leveraging advances in GANs to perform Bayesian inference on high-dimensional simulation-based models.

****

NUQ: Nonparametric Uncertainty Quantification for Deterministic Neural Networks

Nikita Yurevich Kotelevskii,Alexander Fishkov,Kirill Fedyanin,Aleksandr Petiushko,Maxim Panov

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 a variety of real-world image datasets, such as MNIST, SVHN, CIFAR-100 and several versions of ImageNet.

****

Xi-learning: Successor Feature Transfer Learning for General Reward Functions

Chris Reinke,Xavier Alameda-Pineda

Transfer in Reinforcement Learning aims to improve learning performance on target tasks using knowledge from experienced source tasks. Successor features (SF) are a prominent transfer mechanism in domains where the reward function changes between tasks. They reevaluate the expected return of previously learned policies in a new target task and to transfer their knowledge. A limiting factor of the SF framework is its assumption that rewards linearly decompose into successor features and a reward weight vector. We propose a novel SF mechanism, ξ-learning, based on learning the cumulative discounted probability of successor features. Crucially, ξ-learning allows to reevaluate the expected return of policies for general reward functions. We introduce two ξ-learning variations, prove its convergence, and provide a guarantee on its transfer performance. Experimental evaluations based on ξ-learning with function approximation demonstrate the prominent advantage of ξ-learning over available mechanisms not only for general reward functions, but also in the case of linearly decomposable reward functions.

****

Help Me Explore: Minimal Social Interventions for Graph-Based Autotelic Agents

Ahmed Akakzia,Olivier Serris,Olivier Sigaud,Cédric Colas

In the quest for autonomous agents learning open-ended repertoires of skills, most works take a Piagetian perspective: learning trajectories are the results of interactions between developmental agents and their physical environment. The Vygotskian perspective, on the other hand, emphasizes the centrality of the socio-cultural environment: higher cognitive functions emerge from transmissions of socio-cultural processes internalized by the agent. This paper acknowledges these two perspectives and presents GANGSTR, a hybrid agent engaging in both individual and social goal-directed exploration. In a 5-block manipulation domain, GANGSTR discovers and learns to master tens of thousands of configurations. In individual phases, the agent engages in autotelic learning; it generates, pursues and makes progress towards its own goals. To this end, it builds a graph to represent the network of discovered configuration and to navigate between them. In social phases, a simulated social partner suggests goal configurations at the frontier of the agent’s current capabilities. This paper makes two contributions: 1) a m inimal social interaction protocol called Help Me Explore (HME); 2) GANGSTR, a graph-based autotelic agent. As this paper shows, coupling individual and social exploration enables the GANGSTR agent to discover and master the most complex configurations (e.g. stacks of 5 blocks) with only minimal intervention.

****

Privacy Auditing of Machine Learning using Membership Inference Attacks
Jiayuan Ye,Aadyaa Maddi,Sasi Kumar Murakonda,Reza Shokri
Membership inference attacks determine if a given data point is used for training a target model. Thus, this attack could be used as an auditing tool to quantify the private information that a model leaks about the individual data points in its training set. In the last five years, a variety of membership inference attacks against machine learning models are proposed, where each attack exploits a slightly different clue. Also, the attacks are designed under different implicit assumptions about the uncertainties that an attacker has to resolve. Thus attack success rates do not precisely capture the information leakage of models about their data, as they also reflect other uncertainties that the attack algorithm has (for example, about data distribution or characteristics of the target model). In this paper, we present a framework that can explain the implicit assumptions and also the simplifications made in the prior work. We also derive new attack algorithms from our framework that can achieve a high AUC score while also highlighting the different factors that affect their performance. Thus, our algorithms can be used as a tool to perform an accurate and informed estimation of privacy risk in machine learning models. We provide a thorough empirical evaluation of our attack strategies on various machine learning tasks trained on benchmark datasets.

****

Generalization of GANs and overparameterized models under Lipschitz continuity
Khoat Than,Nghia Vu
Generative adversarial networks (GANs) are really complex, and little has been known about their generalization. The existing learning theories lack efficient tools to analyze generalization of GANs. To fill this gap, we introduce a novel tool to analyze generalization: Lipschitz continuity. We demonstrate its simplicity by showing generalization and consistency of overparameterized neural networks. We then use this tool to derive Lipschitz-based generalization bounds for GANs. In particular, our bounds show that penalizing the zero- and first-order informations of the GAN loss will improve generalization. Therefore, this work provides a unified theory for answering the long mystery of why imposing a Lipschitz constraint can help GANs to generalize well in practice.

****

Domain Adversarial Training: A Game Perspective
David Acuna,Marc T Law,Guojun Zhang,Sanja Fidler
The dominant line of work in domain adaptation has focused on learning invariant representations using domain-adversarial training. In this paper, we interpret this approach from a game theoretical perspective. Defining optimal solutions in domain-adversarial training as a local Nash equilibrium, we show that gradient descent in domain-adversarial training can violate the asymptotic convergence guarantees of the optimizer, oftentimes hindering the transfer performance. Our analysis leads us to replace gradient descent with high-order ODE solvers (i.e., Runge–Kutta), for which we derive asymptotic convergence guarantees. This family of optimizers is significantly more stable and allows more aggressive learning rates, leading to high performance gains when used as a drop-in replacement over standard optimizers. Our experiments show that in conjunction with state-of-the-art domain-adversarial methods, we achieve up to 3.5% improvement with less than of half training iterations. Our optimizers are easy to implement, free of additional parameters, and can be plugged into any domain-adversarial framework.

****

AutoMO-Mixer: An automated multi-objective multi-layer perspecton Mixer model for medical image based diagnosis
Xi Chen,Jiahuan Lv,Xuanqing Mou,Zhiguo Zhou
Medical image based diagnosis is one of the most challenging things which is vit

al to human life. Accurately identifying the patient's status through medical images plays an important role in treatment of diseases. Deep learning has achieved great success in medical image analysis. Particularly, Convolutional neural network CNN) can obtain promising performance by learning the features in a supervised way. However, since there are too many parameters to train, CNN always requires a large scale dataset to feed, while it is very difficult to collect the required amount of patient images for a particular clinical problem. Recently, MLP-Mixer (Mixer) which is developed based multiple layer perceptron (MLP) was proposed, in which the number of training parameters is greatly decreased by removing convolutions in the architecture, while it can achieve the similar performance with CNN. Furthermore, obtaining the balanced outcome between sensitivity and specificity is of great importance in patient's status identification. As such, a new automated multi-objective Mixer (AutoMO-Mixer) model was developed in this study. In AutoMO-Mixer, sensitivity and specificity were considered as the objective functions simultaneously to train the model and a Pareto-optimal Mixer model set can be obtained in the training stage. Additionally, since there are several hyperparameters to train, the Bayesian optimization was introduced. To obtain a more reliable results in testing stage, the final output was obtained by fusing the output probabilities of Pareto optimal models through the evidence reasoning (ER) approach. The experimental study demonstrated that AutoMO-Mixer can obtain better performance compared with Mixer and CNN.

****

Faster Reinforcement Learning with Value Target Lower Bounding

Le Zhao,Wei Xu

We show that an arbitrary lower bound of the optimal value function can be used to improve the Bellman value target during value learning. In the tabular case, value learning under the lower bounded Bellman operator converges to the same optimal value as under the original Bellman operator, at a potentially faster speed. In practice, discounted episodic return from the training experience or discounted goal return from hindsight relabeling can serve as the value lower bound when the environment is deterministic. This is because the empirical episodic return from any state can always be repeated through the same action sequence in a deterministic environment, thus a lower bound of the optimal value from the state. We experiment on Atari games, FetchEnv tasks and a challenging physically simulated car push and reach task. We show that in most cases, simply lower bounding with the discounted episodic return performs at least as well as common baselines such as TD3, SAC and Hindsight Experience Replay (HER). It learns much faster than TD3 or HER on some of the harder continuous control tasks, requiring minimal or no parameter tuning.

****

Hessian-Free High-Resolution Nesterov Acceleration for Sampling

Ruilin Li,Hongyuan Zha,Molei Tao

It is known (Shi et al., 2021) that Nesterov's Accelerated Gradient (NAG) for optimization starts to differ from its continuous time limit (noiseless kinetic Langevin) when its stepsize becomes finite. This work explores the sampling counterpart of this phenonemon and proposes an accelerated-gradient-based MCMC method, based on the optimizer of NAG for strongly convex functions (NAG-SC): we reformulate NAG-SC as a Hessian-Free High-Resolution ODE, change its high-resolution coefficient to a hyperparameter, inject appropriate noise, and discretize the resulting diffusion process. Accelerated sampling enabled by the new hyperparameter is quantified and it is not a false acceleration created by time-rescaling. At continuous-time level, additional acceleration over underdamped Langevin in W_2 distance is proved. At discrete algorithm level, a dedicated discretization is proposed to simulate the Hessian-Free High-Resolution SDE in a cost-efficient manner. For log-strong-concave-and-smooth target measures, the proposed algorithm achieves $\tilde{\mathcal{O}}(\sqrt{d}/\epsilon)$ iteration complexity in W_2 distance, same as underdamped Langevin dynamics, but with a reduced constant. Empirical experiments are conducted to numerically verify our theoretical results.

****

Differentiable Expectation-Maximization for Set Representation Learning

Minyoung Kim

We tackle the set2vec problem, the task of extracting a vector representation from an input set comprised of a variable number of feature vectors. Although recent approaches based on self attention such as (Set)Transformers were very successful due to the capability of capturing complex interaction between set elements, the computational overhead is the well-known downside. The inducing-point attention and the latest optimal transport kernel embedding (OTKE) are promising remedies that attain comparable or better performance with reduced computational cost, by incorporating a fixed number of learnable queries in attention. In this paper we approach the set2vec problem from a completely different perspective. The elements of an input set are considered as i.i.d.~samples from a mixture distribution, and we define our set embedding feed-forward network as the maximum-a-posterior (MAP) estimate of the mixture which is approximately attained by a few Expectation-Maximization (EM) steps. The whole MAP-EM steps are differentiable operations with a fixed number of mixture parameters, allowing efficient auto-diff back-propagation for any given downstream task. Furthermore, the proposed mixture set data fitting framework allows unsupervised set representation learning naturally via marginal likelihood maximization aka the empirical Bayes. Interestingly, we also find that OTKE can be seen as a special case of our framework, specifically a single-step EM with extra balanced assignment constraints on the E-step. Compared to OTKE, our approach provides more flexible set embedding as well as prior-induced model regularization. We evaluate our approach on various tasks demonstrating improved performance over the state-of-the-arts.

****

BANANA: a Benchmark for the Assessment of Neural Architectures for Nucleic Acids

Luca Salvatore Lorello,Andrea Galassi,Paolo Torroni

Machine learning has always played an important role in bioinformatics and recent applications of deep learning have allowed solving a new spectrum of biologically relevant tasks.

However, there is still a gap between the ``mainstream'' AI and the bioinformatics communities. This is partially due to the format of bioinformatics data, which are typically difficult to process and adapt to machine learning tasks without deep domain knowledge.

Moreover, the lack of standardized evaluation methods makes it difficult to rigorously compare different models and assess their true performance.

To help to bridge this gap, and inspired by work such as SuperGLUE and TAPE, we present BANANA, a benchmark consisting of six supervised classification tasks designed to assess language model performance in the DNA and RNA domains. The tasks are defined over three genomics and one transcriptomics languages (human DNA, bacterial 16S gene, nematoda ITS2 gene, human mRNA) and measure a model's ability to perform whole-sequence classification in a variety of setups.

Each task was built from readily available data and is presented in a ready-to-use format, with defined labels, splits, and evaluation metrics.

We use BANANA to test state-of-the-art NLP architectures, such as Transformer-based models, observing that, in general, self-supervised pretraining without external corpora is beneficial in every task.

****

Model-Based Robust Adaptive Semantic Segmentation

Jun Wang,Yiannis Kantaros

Semantic image segmentation enjoys a wide range of applications such as autonomous vehicles and medical imaging while it is typically accomplished by deep neural networks (DNNs). Nevertheless, DNNs are known to be fragile to input perturbations that are adversarially crafted or occur due to natural variations, such as changes in weather or lighting conditions. This issue of lack of robustness prevents the application of learning-based semantic segmentation methods on safety-critical applications. To mitigate this challenge, in this paper, we propose model-based robust adaptive training algorithm (MRTAdapt), a new training algorithm to enhance the robustness of DNN-based semantic segmentation methods against natural variations that leverages model-based robust training algorithms and genera tive adversarial networks. Natural variation effects are minimized from both image and label sides. We provide extensive experimental results on both real-world and synthetic datasets demonstrating that model-based robust adaptive training algorithm outperforms multiple state-of-the-art models under various natural variations.

****

Overcoming The Spectral Bias of Neural Value Approximation

Ge Yang,Anurag Ajay,Pulkit Agrawal

Value approximation using deep neural networks is at the heart of off-policy deep reinforcement learning, and is often the primary module that provides learning signals to the rest of the algorithm. While multi-layer perceptron networks are universal function approximators, recent works in neural kernel regression suggest the presence of a \textit{spectral bias}, where fitting high-frequency components of the value function requires exponentially more gradient update steps than the low-frequency ones. In this work, we re-examine off-policy reinforcement learning through the lens of kernel regression and propose to overcome such bias via a composite neural tangent kernel. With just a single line-change, our approach, the Fourier feature networks (FFN) produce state-of-the-art performance on challenging continuous control domains with only a fraction of the compute. Faster convergence and better off-policy stability also make it possible to remove the target network without suffering catastrophic divergences, which further reduces TD(0)'s estimation bias on a few tasks. Code and analysis available at https://geyang.github.io/ffn.

****

Improving zero-shot generalization in offline reinforcement learning using generalized similarity functions

Bogdan Mazoure,Ilya Kostrikov,Ofir Nachum,Jonathan Tompson

Reinforcement learning (RL) agents are widely used for solving complex sequential decision making tasks, but still exhibit difficulty in 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 also improving zero-shot generalization performance on a complex offline RL benchmark, offline Procgen.

****

Sample-efficient actor-critic algorithms with an etiquette for zero-sum Markov games

Ahmet Alacaoglu,Luca Viano,Niao He,Volkan Cevher

We introduce algorithms based on natural policy gradient and two time-scale natural actor-critic, and analyze their sample complexity for solving two player zero-sum Markov games in the tabular case. Our results improve the best-known sample complexities of policy gradient/actor-critic methods for convergence to Nash equilibrium in the multi-agent setting. We use the error propagation scheme in approximate dynamic programming, recent advances for global convergence of policy gradient methods, temporal difference learning, and techniques from stochastic primal-dual optimization literature. Our algorithms feature two stages, requiring agents to agree on an etiquette before starting their interactions, which is feasible for instance in self-play. On the other hand, the agents only access to joint reward and joint next state and not to each other's actions or policies. Our sample complexities also match the best-known results for global convergence of policy gradient and two time-scale actor-critic algorithms in the single agent setting. We provide numerical verification of our method for a two-player bandi t environment and a two player game, Alesia. We observe improved empirical performance as compared to the recently proposed optimistic gradient descent ascent variant for Markov games.

****

Towards the Memorization Effect of Neural Networks in Adversarial Training

Han Xu,Xiaorui Liu,Wentao Wang,Wenbiao Ding,Zhongqin Wu,Zitao Liu,Anil Jain,Jiliang Tang

Recent studies suggest that "memorization" is one important factor for overparameterized deep neural networks (DNNs) to achieve optimal performance. Specifically, the perfectly fitted DNNs can memorize the labels of many atypical samples, generalize their memorization to correctly classify test atypical samples and enjoy better test performance. While, DNNs which are optimized via adversarial training algorithms can also achieve perfect training performance by memorizing the labels of atypical samples, as well as the adversarially perturbed atypical samples. However, adversarially trained models always suffer from poor generalization, with both relatively low clean accuracy and robustness on the test set. In this work, we study the effect of memorization in adversarial trained DNNs and disclose two important findings: (a) Memorizing atypical samples is only effective to improve DNN's accuracy on clean atypical samples, but hardly improve their adversarial robustness and (b) Memorizing certain atypical samples will even hurt the DNN's performance on typical samples. Based on these two findings, we propose Benign Adversarial Training (BAT) which can facilitate adversarial training to avoid fitting "harmful" atypical samples and fit as more "benign" atypical samples as possible. In our experiments, we validate the effectiveness of BAT, and show it can achieve better clean accuracy vs. robustness trade-off than baseline methods, in benchmark datasets such as CIFAR100 and Tiny ImageNet.

****

Match Prediction Using Learned History Embeddings

Maxwell Goldstein,Leon Bottou,Rob Fergus

Contemporary ranking systems that are based on win/loss history, such as Elo or TrueSkill represent each player using a scalar estimate of ability (plus variance, in the latter case). While easily interpretable, this approach has a number of shortcomings: (i) latent attributes of a player cannot be represented, and (ii) it cannot seamlessly incorporate contextual information (e.g. home-field advantage). In this work, we propose a simple Transformer-based approach for pairwise competitions that recursively operates on game histories, rather than modeling players directly. By characterizing each player entirely by its history, rather than an underlying scalar skill estimate, it is able to make accurate predictions even for new players with limited history. Additionally, it is able to model both transitive and non-transitive relations and can leverage contextual information. When restricted to the same information as Elo and Glicko, our approach significantly outperforms them on predicting the outcome of real-world Chess, Baseball and Ice Hockey games. %Further gains can be achieved when game meta-data is added.

****

Invariance Through Inference

Takuma Yoneda,Ge Yang,Matthew Walter,Bradly C. Stadie

We introduce a general approach, called invariance through inference, for improving the test-time performance of a behavior agent in deployment environments with unknown perceptual variations. Instead of producing invariant visual features through memorization, invariance through inference turns adaptation at deployment-time into an unsupervised learning problem by trying to match the distribution of latent features to the agent's prior experience without relying on paired data. Although simple, we show that this idea leads to surprising improvements on a variety of adaptation scenarios without access to task reward, including changes in camera poses from the challenging distractor control suite.

****

GNN-LM: Language Modeling based on Global Contexts via GNN

Yuxian Meng,Shi Zong,Xiaoya Li,Xiaofei Sun,Tianwei Zhang,Fei Wu,Jiwei Li

Inspired by the notion that "it to copy is easier than to memorize", in this work, we introduce GNN-LM, which extends vanilla neural language model (LM) by allowing to reference similar contexts in the entire training corpus. We build a directed heterogeneous graph between an input context and its semantically related neighbors selected from the training corpus, where nodes are tokens in the input context and retrieved neighbor contexts, and edges represent connections between nodes. Graph neural networks (GNNs) are constructed upon the graph to aggregate information from similar contexts to decode the token. This learning paradigm provides direct access to the reference contexts and helps improve a model's generalization ability. We conduct comprehensive experiments to validate the effectiveness of the GNN-LM: GNN-LM achieves a new state-of-the-art perplexity of 14.8 on WikiText-103 (a 3.9 point improvement over its counterpart of the vanilla LM model), and shows substantial improvement on One Billion Word and Enwiki8 datasets against strong baselines. In-depth ablation studies are performed to understand the mechanics of GNN-LM. The code can be found at https://github.com/ShannonAI/GNN-LM.

****

Invariant Causal Mechanisms through Distribution Matching

Mathieu Chevalley,Charlotte Bunne,Andreas Krause,Stefan Bauer

Learning representations that capture the underlying data generating process is akey problem for data efficient and robust use of neural networks. One key property for robustness which the learned representation should capture and which recently received a lot of attention is described by the notion of invariance. In this work we provide a causal perspective and new algorithm for learning invariant representations. Empirically we show that this algorithm works well on a diverse set of tasks and in particular we observe state-of-the-art performance on domain generalization, where we are able to significantly boost the score of existing models.

****

Prospect Pruning: Finding Trainable Weights at Initialization using Meta-Gradients

Milad Alizadeh,Shyam A. Tailor,Luisa M Zintgraf,Joost van Amersfoort,Sebastian Farquhar,Nicholas Donald Lane,Yarin Gal

Pruning neural networks at initialization would enable us to find sparse models that retain the accuracy of the original network while consuming fewer computational resources for training and inference. However, current methods are insufficient to enable this optimization and lead to a large degradation in model performance. In this paper, we identify a fundamental limitation in the formulation of current methods, namely that their saliency criteria look at a single step at the start of training without taking into account the trainability of the network. While pruning iteratively and gradually has been shown to improve pruning performance, explicit consideration of the training stage that will immediately follow pruning has so far been absent from the computation of the saliency criterion. To overcome the short-sightedness of existing methods, we propose Prospect Pruning (ProsPr), which uses meta-gradients through the first few steps of optimization to determine which weights to prune. ProsPr combines an estimate of the higher-order effects of pruning on the loss and the optimization trajectory to identify the trainable sub-network. Our method achieves state-of-the-art pruning performance on a variety of vision classification tasks, with less data and in a single shot compared to existing pruning-at-initialization methods.

****

Compound Multi-branch Feature Fusion for Real Image Restoration

Chi-Mao Fan,Tsung-Jung Liu,Kuan-Hsien Liu

Image restoration is a challenging and ill-posed problem which also has been a long-standing issue. However, most of learning based restoration methods are proposed to target one degradation type which means they are lack of generalization. In this paper, we proposed a multi-branch restoration model inspired from the Human Visual System (i.e., Retinal Ganglion Cells) which can achieve multiple restoration tasks in a general framework. The experiments show that the proposed multi-branch architecture, called CMFNet, has competitive performance results on f our datasets, including image dehazing, deraindrop, and deblurring, which are very common applications for autonomous cars. The source code and pretrained models of three restoration tasks are available at https://github.com/publish_after_accepting/CMFNet.

****

An Optimization Perspective on Realizing Backdoor Injection Attacks on Deep Neural Networks in Hardware

M. Caner Tol,Saad Islam,Berk Sunar,Ziming Zhang

State-of-the-art deep neural networks (DNNs) have been proven to be vulnerable to adversarial manipulation and backdoor attacks. Backdoored models deviate from expected behavior on inputs with predefined triggers while retaining performance on clean data. Recent works focus on software simulation of backdoor injection during the inference phase by modifying network weights, which we find often unrealistic in practice due to the hardware restriction such as bit allocation in memory. In contrast, in this work, we investigate the viability of backdoor injection attacks in real-life deployments of DNNs on hardware and address such practical issues in hardware implementation from a novel optimization perspective.

We are motivated by the fact that the vulnerable memory locations are very rare, device-specific, and sparsely distributed. Consequently, we propose a novel network training algorithm based on constrained optimization for realistic backdoor injection attack in hardware. By modifying parameters uniformly across the convolutional and fully-connected layers as well as optimizing the trigger pattern together, we achieve the state-of-the-art attack performance with fewer bit flips. For instance, our method on a hardware-deployed ResNet-20 model trained on CIFAR-10 can achieve over 91\% test accuracy and 94\% attack success rate by flipping only 10 bits out of 2.2 million bits.

****

Learn the Time to Learn: Replay Scheduling for Continual Learning

Marcus Klasson,Hedvig Kjellstrom,Cheng Zhang

Replay-based continual learning has shown to be successful in mitigating catastrophic forgetting. Most previous works focus on increasing the sample quality in the commonly small replay memory. However, in many real-world applications, replay memories would be limited by constraints on processing time rather than storage capacity as most organizations do store all historical data in the cloud. Inspired by human learning, we illustrate that scheduling over which tasks to revisit is critical to the final performance with finite memory resources. To this end, we propose to learn the time to learn for a continual learning system, in which we learn schedules over which tasks to replay at different times using Monte Carlo tree search. We perform extensive evaluation and show that our method can learn replay schedules that significantly improve final performance across all tasks than baselines without considering the scheduling. Furthermore, our method can be combined with any other memory selection methods leading to consistently improved performance. Our results indicate that the learned schedules are also consistent with human learning insights.

****

Foreground-attention in neural decoding: Guiding Loop-Enc-Dec to reconstruct visual stimulus images from fMRI

Kai Chen,Yongqiang Ma,Mingyang Sheng,Nanning Zheng

The reconstruction of visual stimulus images from functional Magnetic Resonance Imaging (fMRI) has received extensive attention in recent years, which provides a possibility to interpret the human brain. Due to the high-dimensional and high-noise characteristics of fMRI data, how to extract stable, reliable and useful information from fMRI data for image reconstruction has become a challenging problem. Inspired by the mechanism of human visual attention, in this paper, we propose a novel method of reconstructing visual stimulus images, which first decodes the distribution of visual attention from fMRI, and then reconstructs the visual images guided by visual attention. We define visual attention as foreground attention (F-attention). Because the human brain is strongly wound into sulci and gyri, some spatially adjacent voxels are not connected in practice. Therefore, it is necessary to consider the global information when decoding fMRI, so we int roduce the self-attention module for capturing global information into the process of decoding F-attention. In addition, in order to obtain more loss constraints in the training process of encoder-decoder, we also propose a new training strategy called Loop-Enc-Dec. The experimental results show that the F-attention decoder decodes the visual attention from fMRI successfully, and the Loop-Enc-Dec guided by F-attention can also well reconstruct the visual stimulus images.

****

CoMPS: Continual Meta Policy Search

Glen Berseth,Zhiwei Zhang,Grace Zhang,Chelsea Finn,Sergey Levine

We develop a new continual meta-learning method to address challenges in sequential multi-task learning. In this setting, the agent's goal is to achieve high reward over any sequence of tasks quickly. Prior meta-reinforcement learning algorithms have demonstrated promising results in accelerating the acquisition of new tasks. However, they require access to all tasks during training. Beyond simply transferring past experience to new tasks, our goal is to devise continual reinforcement learning algorithms that learn to learn, using their experience on previous tasks to learn new tasks more quickly. We introduce a new method, continual meta-policy search (CoMPS), that removes this limitation by meta-training in an incremental fashion, over each task in a sequence, without revisiting prior tasks. CoMPS continuously repeats two subroutines: learning a new task using RL and using the experience from RL to perform completely offline meta-learning to prepare for subsequent task learning. We find that CoMPS outperforms prior continual learning and off-policy meta-reinforcement methods on several sequences of challenging continuous control tasks.

****

Soteria: In search of efficient neural networks for private inference

Anshul Aggarwal,Trevor E Carlson,Reza Shokri,Shruti Tople

In the context of ML as a service, our objective is to protect the confidentiality of the users’ queries and the server's model parameters, with modest computation and communication overhead. Prior solutions primarily propose fine-tuning cryptographic methods to make them efficient for known fixed model architectures. The drawback with this line of approach is that the model itself is never designed to efficiently operate with existing cryptographic computations. We observe that the network architecture, internal functions, and parameters of a model, which are all chosen during training, significantly influence the computation and communication overhead of a cryptographic method, during inference.Thus, we propose SOTERIA — a training method to construct model architectures that are by-design efficient for private inference. We use neural architecture search algorithms with the dual objective of optimizing the accuracy of the model and the overhead of using cryptographic primitives for secure inference. Given the flexibility of modifying a model during training, we find accurate models that are also efficient for private computation. We select garbled circuits as our underlying cryptographic primitive, due to their expressiveness and efficiency. We empirically evaluate SOTERIA on MNIST and CIFAR10 datasets, to compare with the prior work on secure inference. Our results confirm that SOTERIA is indeed effective in balancing performance and accuracy.

****

Orthogonalising gradients to speedup neural network optimisation

Mark Tuddenham,Adam Prugel-Bennett,Jonathon Hare

The optimisation of neural networks can be sped up by orthogonalising the gradients before the optimisation step, ensuring the diversification of the learned representations. We hypothesize that components in the same layer learn the same representations at the beginning of learning. To prevent this we orthogonalise the gradients of the components with respect to each other.

Our method of orthogonalisation allows the weights to be used more flexibly, in contrast to restricting the weights to an orthogonalised sub-space. We tested this method on ImageNet and CIFAR-10 resulting in a large decrease in learning time, and also obtain a speed-up on the semi-supervised learning BarlowTwins. We obtain similar accuracy to SGD without fine-tuning and better accuracy for naïvely chosen hyper-parameters.

****

Non-Parametric Neuro-Adaptive Control Subject to Task Specifications
Christos Verginis,Zhe Xu,ufuk topcu

We develop a learning-based algorithm for the control of autonomous systems governed by unknown, nonlinear dynamics to satisfy user-specified spatio-temporal tasks expressed as signal temporal logic specifications. Most existing algorithms either assume certain parametric forms for the unknown dynamic terms or resort to unnecessarily large control inputs in order to provide theoretical guarantees.

The proposed algorithm addresses these drawbacks by integrating neural-network-based learning with adaptive control. More specifically, the algorithm learns a controller, represented as a neural network, using training data that correspond to a collection of system parameters and tasks. These parameters and tasks are derived by varying the nominal parameters and the spatio-temporal constraints of the user-specified task, respectively. It then incorporates this neural network into an online closed-form adaptive control policy in such a way that the resulting behavior satisfies the user-defined task. The proposed algorithm does not use any a priori information on the unknown dynamic terms or any approximation schemes. We provide formal theoretical guarantees on the satisfaction of the task. Numerical experiments on a robotic manipulator and a unicycle robot demonstrate that the proposed algorithm guarantees the satisfaction of 50 user-defined tasks, and outperforms control policies that do not employ online adaptation or the neural-network controller. Finally, we show that the proposed algorithm achieves greater performance than standard reinforcement-learning algorithms in the pendulum benchmarking environment.

****

Accelerating Optimization using Neural Reparametrization
Nima Dehmamy,Csaba Both,Jianzhi Long,Rose Yu

We tackle the problem of accelerating certain optimization problems related to steady states in ODE and energy minimization problems common in physics.
We reparametrize the optimization variables as the output of a neural network.
We then find the conditions under which this neural reparameterization could speed up convergence rates during gradient descent.
We find that to get the maximum speed up the neural network needs to be a special graph convolutional network (GCN) with its aggregation function constructed from the gradients of the loss function.
We show the utility of our method on two different optimization problems on graphs and point-clouds.

****

Losing Less: A Loss for Differentially Private Deep Learning
Ali Shahin Shamsabadi,Nicolas Papernot

Differentially Private Stochastic Gradient Descent, DP-SGD, is the canonical approach to training deep neural networks with guarantees of Differential Privacy (DP). However, the modifications DP-SGD introduces to vanilla gradient descent negatively impact the accuracy of deep neural networks. In this paper, we are the first to observe that some of this performance can be recovered when training with a loss tailored to DP-SGD; we challenge cross-entropy as the de facto loss for deep learning with DP. Specifically, we introduce a loss combining three terms: the summed squared error, the focal loss, and a regularization penalty. The first term encourages learning with faster convergence. The second term emphasizes hard-to-learn examples in the later stages of training. Both are beneficial because the privacy cost of learning increases with every step of DP-SGD. The third term helps control the sensitivity of learning, decreasing the bias introduced by gradient clipping in DP-SGD. Using our loss function, we achieve new state-of-the-art tradeoffs between privacy and accuracy on MNIST, FashionMNIST, and CIFAR10. Most importantly, we improve the accuracy of DP-SGD on CIFAR10 by 4% for a DP guarantee of $\varepsilon=3$.

****

Teamwork makes von Neumann work:Min-Max Optimization in Two-Team Zero-Sum Games

Fivos Kalogiannis,Ioannis Panageas,Emmanouil-Vasileios Vlatakis-Gkaragkounis

Motivated by recent advances in both theoretical and applied aspects of multiplayer games, spanning from e-sports to multi-agent generative adversarial networks, we focus on min-max optimization in team zero-sum games. In this class of games, players are split in two teams with payoffs equal within the same team and of opposite sign across the opponent team. Unlike the textbook two player zero-sum games, finding a Nash equilibrium in our class can be shown to be $\textsf{CLS}$-hard, i.e., it is unlikely to have a polynomial time algorithm for computing Nash equilibria. Moreover In this generalized framework, we establish that even asymptotic last iterate or time average convergence to a Nash Equilibrium is not possible using Gradient Descent Ascent (GDA), its optimistic variant and extra gradient. Specifically, we present a family of team games whose induced utility is non-multilinear with non-attractive $\textit{per-se}$ mixed Nash Equilibria, as strict saddle points of the underlying optimization landscape. Leveraging techniques from control theory, we complement these negative results by designing a modified GDA that converges locally to Nash equilibria. Finally, we discuss connections of our framework with AI architectures with team competition structure like multi-agent generative adversarial networks.

****

Generalized rectifier wavelet covariance models for texture synthesis

Antoine Brochard,Sixin Zhang,Stéphane Mallat

State-of-the-art maximum entropy models for texture synthesis are built from statistics relying on image representations defined by convolutional neural networks (CNN). Such representations capture rich structures in texture images, outperforming wavelet-based representations in this regard. However, conversely to neural networks, wavelets offer meaningful representations, as they are known to detect structures at multiple scales (e.g. edges) in images. In this work, we propose a family of statistics built upon non-linear wavelet based representations, that can be viewed as a particular instance of a one-layer CNN, using a generalized rectifier non-linearity. These statistics significantly improve the visual quality of previous classical wavelet-based models, and allow one to produce syntheses of similar quality to state-of-the-art models, on both gray-scale and color textures. We further provide insights on memorization effects in these models.

****

Learning Graph Representations for Influence Maximization

George Panagopoulos,Nikolaos Tziortziotis,Fragkiskos D. Malliaros,Michalis Vazirgiannis

As the field of machine learning for combinatorial optimization advances, traditional problems are resurfaced and readdressed through this new perspective. The overwhelming majority of the literature focuses on small graph problems, while several real-world problems are devoted to large graphs. Here, we focus on two such problems: influence estimation, a #P-hard counting problem, and influence maximization, an NP-hard problem. We develop Glie, a Graph Neural Network (GNN) that inherently parameterizes an upper bound of influence estimation and train it on small simulated graphs. Experiments show that Glie provides accurate influence estimation for real graphs up to 10 times larger than the train set. More importantly, it can be used for influence maximization on considerably larger graphs, as the predictions ranking is not affected by the drop of accuracy. We develop a version of Cost Effective Lazy Forward optimization with Glie instead of simulated influence estimation, surpassing the benchmark for influence maximization, although with a computational overhead. To balance the time complexity and quality of influence, we propose two different approaches. The first is a Q-network that learns to choose seeds sequentially using Glie's predictions. The second defines a provably submodular function based on Glie's representations to rank nodes fast while building the seed set. The latter provides the best combination of time efficiency and influence spread, outperforming SOTA benchmarks.

****

Iterative Decoding for Compositional Generalization in Transformers

Luana Ruiz,Joshua Ainslie,Santiago Ontanon

Deep learning models do well at generalizing to in-distribution data but struggle to generalize compositionally, i.e., to combine a set of learned primitives to solve more complex tasks. In particular, in sequence-to-sequence (seq2seq) learning, transformers are often unable to predict even marginally longer examples than those seen during training. This paper introduces iterative decoding, an alternative to seq2seq learning that (i) improves transformer compositional generalization and (ii) evidences that, in general, seq2seq transformers do not learn iterations that are not unrolled. Inspired by the idea of compositionality---that complex tasks can be solved by composing basic primitives---training examples are broken down into a sequence of intermediate steps that the transformer then learns iteratively. At inference time, the intermediate outputs are fed back to the transformer as intermediate inputs until an end-of-iteration token is predicted. Through numerical experiments, we show that transfomers trained via iterative decoding outperform their seq2seq counterparts on the PCFG dataset, and solve the problem of calculating Cartesian products between vectors longer than those seen during training with 100% accuracy, a task at which seq2seq models have been shown to fail. We also illustrate a limitation of iterative decoding, specifically, that it can make sorting harder to learn on the CFQ dataset.

****

The Low-Rank Simplicity Bias in Deep Networks
Minyoung Huh,Hossein Mobahi,Richard Zhang,Brian Cheung,Pulkit Agrawal,Phillip Isola

Modern deep neural networks are highly over-parameterized compared to the data on which they are trained, yet they often generalize remarkably well. A flurry of recent work has asked: why do deep networks not overfit to their training data? In this work, we make a series of empirical observations that investigate the hypothesis that deeper networks are inductively biased to find solutions with lower rank embeddings. We conjecture that this bias exists because the volume of functions that maps to low-rank embedding increases with depth. We show empirically that our claim holds true on finite width linear and non-linear models and show that these are the solutions that generalize well. We then show that the low-rank simplicity bias exists even after training, using a wide variety of commonly used optimizers. We found this phenomenon to be resilient to initialization, hyper-parameters, and learning methods. We further demonstrate how linear over-parameterization of deep non-linear models can be used to induce low-rank bias, improving generalization performance without changing the effective model capacity. Practically, we demonstrate that simply linearly over-parameterizing standard models at training time can improve performance on image classification tasks, including ImageNet.

****

Towards Evaluating the Robustness of Neural Networks Learned by Transduction
Jiefeng Chen,Xi Wu,Yang Guo,Yingyu Liang,Somesh Jha

There has been emerging interest in using transductive learning for adversarial robustness (Goldwasser et al., NeurIPS 2020; Wu et al., ICML 2020; Wang et al., ArXiv 2021). Compared to traditional defenses, these defense mechanisms "dynamically learn" the model based on test-time input; and theoretically, attacking these defenses reduces to solving a bilevel optimization problem, which poses difficulty in crafting adaptive attacks. In this paper, we examine these defense mechanisms from a principled threat analysis perspective. We formulate and analyze threat models for transductive-learning based defenses, and point out important subtleties. We propose the principle of attacking model space for solving bilevel attack objectives, and present Greedy Model Space Attack (GMSA), an attack framework that can serve as a new baseline for evaluating transductive-learning based defenses. Through systematic evaluation, we show that GMSA, even with weak instantiations, can break previous transductive-learning based defenses, which were resilient to previous attacks, such as AutoAttack (Croce and Hein, ICML 2020). On the positive side, we report a somewhat surprising empirical result of "transductive adversarial training": Adversarially retraining the model using fresh randomness at the test time gives a significant increase in robustness against attacks we consider.

****

Evaluating generative networks using Gaussian mixtures of image features

Lorenzo Luzi,Carlos Ortiz Marrero,Nile N Wynar,Richard Baraniuk,Michael J. Henry

We develop a measure for evaluating the performance of generative networks given two sets of images. A popular performance measure currently used to do this is the Fréchet Inception Distance (FID). However, FID assumes that images featurized using the penultimate layer of Inception-v3 follow a Gaussian distribution. This assumption allows FID to be easily computed, since FID uses the 2-Wasserstein distance of two Gaussian distributions fitted to the featurized images. However, we show that Inception-v3 features of the ImageNet dataset are not Gaussian; in particular, each marginal is not Gaussian. To remedy this problem, we model the featurized images using Gaussian mixture models (GMMs) and compute the 2-Wasserstein distance restricted to GMMs. We define a performance measure, which we call WaM, on two sets of images by using Inception-v3 (or another classifier) to featurize the images, estimate two GMMs, and use the restricted 2-Wasserstein distance to compare the GMMs. We experimentally show the advantages of WaM over FID, including how FID is more sensitive than WaM to image perturbations. By modelling the non-Gaussian features obtained from Inception-v3 as GMMs and using a GMM metric, we can more accurately evaluate generative network performance.

****

OBJECT DYNAMICS DISTILLATION FOR SCENE DECOMPOSITION AND REPRESENTATION

Qu Tang,Xiangyu Zhu,Zhen Lei,Zhaoxiang Zhang

The ability to perceive scenes in terms of abstract entities is crucial for us to
achieve higher-level intelligence. Recently, several methods have been proposed to learn object-centric representations of scenes with multiple objects, yet most
of which focus on static scenes. In this paper, we work on object dynamics and propose Object Dynamics Distillation Network (ODDN), a framework that distillates explicit object dynamics (e.g., velocity) from sequential static representations. ODDN also builds a relation module to model object interactions. We verify our approach on tasks of video reasoning and video prediction, which are two important evaluations for video understanding. The results show that the reasoning model with visual representations of ODDN performs better in answering reasoning questions around physical events in a video compared to the previous state-of-the-art methods. The distilled object dynamics also could be used to predict future video frames given two input frames, involving occlusion and objects collision. In addition, our architecture brings better segmentation quality and higher
reconstruction accuracy.

****

Practical Integration via Separable Bijective Networks

Christopher M Bender,Patrick Emmanuel,Michael K. Reiter,Junier Oliva

Neural networks have enabled learning over examples that contain thousands of dimensions.
However, most of these models are limited to training and evaluating on a finite collection of \textit{points} and do not consider the hypervolume in which the data resides.
Any analysis of the model's local or global behavior is therefore limited to very expensive or imprecise estimators.
We propose to formulate neural networks as a composition of a bijective (flow) network followed by a learnable, separable network.
This construction allows for learning (or assessing) over full hypervolumes with precise estimators at tractable computational cost via integration over the \textit{input space}.
We develop the necessary machinery, propose several practical integrals to use during training, and demonstrate their utility.

****

CONTROLLING THE MEMORABILITY OF REAL AND UNREAL FACE IMAGES

Mohammad Younesi,Yalda Mohsenzadeh

Every day, we are bombarded with many face photographs, whether on social media, television, or smartphones. From an evolutionary perspective, faces are intended to be remembered, mainly due to survival and personal relevance. However, all these faces do not have the equal opportunity to stick in our minds. It has been shown that memorability is an intrinsic feature of an image but yet, it’s largely unknown what attributes make the images more memorable. In this work, we aim to address this question by proposing a fast approach to modify and control the memorability of face images. In our proposed method, we first find a hyperplane in the latent space of StyleGAN to separate high and low memorable images. We then modify the image memorability (while keeping the identity and other facial features such as age, emotion, etc.) by moving in the positive or negative direction of this hyperplane normal vector. We further analyzed how different layers of the styleGAN augmented latent space contribute to face memorability. These analyses showed how each individual face attribute makes images more or less memorable. Most importantly, we evaluated our proposed method for both real and unreal (generated) face images. The proposed method successfully modifies and controls the memorability of real human faces as well as unreal(generated) faces. Our proposed method can be employed in photograph editing applications for social media, learning aids, or advertisement purposes.

****

Regularized-OFU: an efficient algorithm for general contextual bandit with optimization oracles

Yichi Zhou,Shihong Song,Huishuai Zhang,Jun Zhu,Wei Chen,Tie-Yan Liu

In contextual bandit, one major challenge is to develop theoretically solid and empirically efficient algorithms for general function classes. We present a novel algorithm called \emph{regularized optimism in face of uncertainty (ROFU)} for general contextual bandit problems. It exploits an optimization oracle to calculate the well-founded upper confidence bound (UCB). Theoretically, for general function classes under very mild assumptions, it achieves a near-optimal regret bound $\Tilde{O}(\sqrt{T})$. Practically, one great advantage of ROFU is that the optimization oracle can be efficiently implemented with low computational cost. Thus, we can easily extend ROFU for contextual bandits with deep neural networks as the function class, which outperforms strong baselines including the UCB and Thompson sampling variants.

****

Self-Joint Supervised Learning

Navid Kardan,Mubarak Shah,Mitch Hill

Supervised learning is a fundamental framework used to train machine learning systems. A supervised learning problem is often formulated using an i.i.d. assumption that restricts model attention to a single relevant signal at a time when predicting. This contrasts with the human ability to actively use related samples as reference when making decisions. We hypothesize that the restriction to a single signal for each prediction in the standard i.i.d. framework contributes to well-known drawbacks of supervised learning: making overconfident predictions and vulnerability to overfitting, adversarial attacks, and out-of-distribution data. To address these limitations, we propose a new supervised learning paradigm called self-joint learning that generalizes the standard approach by modeling the joint conditional distribution of two observed samples, where each sample is an image and its label. Rather than assuming samples are independent, our models explicitly learn the sample-to-sample relation of conditional independence. Our framework can naturally incorporate auxiliary unlabeled data to further improve the performance. Experiments on benchmark image datasets show our method offers significant improvement over standard supervised learning in terms of accuracy, robustness against adversarial attacks, out-of-distribution detection, and overconfidence mitigation.

****

Fast and Sample-Efficient Domain Adaptation for Autoencoder-Based End-to-End Communication

Jayaram Raghuram,Yijing Zeng,Dolores Garcia,Somesh Jha,Suman Banerjee,Joerg Widmer,Rafael Ruiz

The problem of domain adaptation conventionally considers the setting where a source domain has plenty of labeled data, and a target domain (with a different data distribution) has plenty of unlabeled data but none or very limited labeled data. In this paper, we address the setting where the target domain has only limited labeled data from a distribution that is expected to change frequently. We first propose a fast and light-weight method for adapting a Gaussian mixture density network (MDN) using only a small set of target domain samples. This method is well-suited for the setting where the distribution of target data changes rapidly (e.g., a wireless channel), making it challenging to collect a large number of samples and retrain. We then apply the proposed MDN adaptation method to the problem of end-of-end learning of a communication autoencoder, which jointly learns the encoder, decoder, and a channel networks to minimize the decoding error rate. However, the error rate of an autoencoder trained on a particular (source) channel distribution can degrade as the channel distribution changes frequently, not allowing enough time for data collection and retraining of the autoencoder to the target channel distribution. We propose a method for adapting the autoencoder without modifying the encoder and decoder neural networks, and adapting only the MDN model of the channel. The method utilizes feature transformations at the decoder to compensate for changes in the channel distribution, and effectively present to the decoder samples close to the source distribution. Experimental evaluation on simulated datasets and real mmWave wireless channels demonstrate that the proposed methods can adapt the MDN model using very limited number of samples, and improve or maintain the error rate of the autoencoder under changing channel conditions.

****

Mistake-driven Image Classification with FastGAN and SpinalNet

Mohit Kumar Ahuja,Sahil Sahil,Helge Spieker

Image classification with classes of varying difficulty can cause performance disparity in deep learning models and reduce the overall performance and reliability of the predictions. In this paper, we address the problem of imbalanced performance in image classification, where the trained model has performance deficits in some of the dataset's classes. By employing Generative Adversarial Networks (GANs) to augment these deficit classes, we finetune the model towards a balanced performance among the different classes and an overall better performance on the whole dataset. Specifically, we combine a light-weight GAN method, FastGAN (Liu et al., 2021), for class-wise data augmentation with Progressive SpinalNet (Chopra, 2021) and Sharpness-Aware Minimization (SAM) (Foret et al., 2020) for training. Unlike earlier works, during training, our method focuses on those classes with lowest accuracy after the initial training phase, which leads to better performance. Only these classes are augmented to boost the accuracy. Due to the use of a light-weight GAN method, the GAN-based augmentation is viable and effective for mistake-driven training even for datasets with only few images per class, while simultaneously requiring less computation than other, more complex GAN methods. Our extensive experiments, including ablation studies on all key components, show competitive or better accuracy than the previous state-of-the-art on five datasets with different sizes and image resolutions.

****

Rethinking Supervised Pre-Training for Better Downstream Transferring

Yutong Feng,Jianwen Jiang,Mingqian Tang,Rong Jin,Yue Gao

The pretrain-finetune paradigm has shown outstanding performance on many applications of deep learning, where a model is pre-trained on an upstream large dataset (e.g. ImageNet), and is then fine-tuned to different downstream tasks. Though for most cases, the pre-training stage is conducted based on supervised methods, recent works on self-supervised pre-training have shown powerful transferability and even outperform supervised pre-training on multiple downstream tasks. It thus remains an open question how to better generalize supervised pre- training model to downstream tasks. In this paper, we argue that the worse transferability of existing supervised pre-training methods arise from the negligence of valuable intra-class semantic difference. This is because these methods tend to push images from the same class close to each other despite of the large diversity in their visual contents, a problem to which referred as “overfit of upstream tasks”. To alleviate this problem, we propose a new supervised pre-training method based on Leave-One-Out K-Nearest-Neighbor, or LOOK for short. It relieves the problem of overfitting upstream tasks by only requiring each image to share its class label with most of its k nearest neighbors, thus allowing each class to exhibit a multi-mode distribution and consequentially preserving part of intra-class difference for better transferring to downstream tasks. We developed efficient implementation of the proposed method that scales well to large datasets. Experimental studies on multiple downstream tasks show that LOOK outperforms other state-of-the-art methods for supervised and self-supervised pre-training.

****

Finding One Missing Puzzle of Contextual Word Embedding: Representing Contexts as Manifold

Hailin Hu,Rong Yao,Cheng LI

The current understanding of contextual word embedding interprets the representation by associating each token to a vector that is dynamically modulated by the context. However, this “token-centric” understanding does not explain how a model represents context itself, leading to a lack of characterization from such a perspective. In this work, to establish a rigorous definition of “context representation”, we formalize this intuition using a category theory framework, which indicates the necessity of including the information from both tokens and how transitions happen among different tokens in a given context. As a practical instantiation of our theoretical understanding, we also show how to leverage a manifold learning method to characterize how a representation model (i.e., BERT) encodes different contexts and how a representation of context changes when going through different components such as attention and FFN. We hope this novel theoretic perspective sheds light on the further improvements in Transformer-based language representation models.

****

On the benefits of deep RL in accelerated MRI sampling

Thomas Sanchez,Igor Krawczuk,Volkan Cevher

Deep learning approaches have shown great promise in accelerating magnetic resonance imaging (MRI), by reconstructing high quality images from highly undersampled data. While previous sampling methods relied on heuristics, recent work has improved the state-of-the-art (SotA) with deep reinforcement learning (RL) sampling policies, which promise the possibility of long term planning and adapting to the observations at test time. In this work, we perform a careful reproduction and comparison of SotA RL sampling methods. We find that i) a simple, easy-to-code, greedily trained fixed policy can match or outperform deep RL methods and ii) find and resolve subtle variations in the preprocessing which previously made results incomparable across different works.

Our results cast doubt on the added value of current RL approaches over fixed masks in MRI sampling and highlight the importance of leveraging strong fixed baselines, standardized reporting as well as isolating the source of improvement in a given work via ablations. We conclude with recommendations for the training and evaluation of deep reconstruction and sampling systems for adaptive MRI based on our findings.

****

Automatic Forecasting via Meta-Learning

Mustafa Abdallah,Ryan Rossi,Kanak Mahadik,Sungchul Kim,Handong Zhao,Haoliang Wang,Saurabh Bagchi

In this work, we develop techniques for fast automatic selection of the best forecasting model for a new unseen time-series dataset, without having to first train (or evaluate) all the models on the new time-series data to select the best one. In particular, we develop a forecasting meta-learning approach called AutoForecast that allows for the quick inference of the best time-series forecasting model for an unseen dataset. Our approach learns both forecasting models performances over time horizon of same dataset and task similarity across different datasets. The experiments demonstrate the effectiveness of the approach over state-of-the-art (SOTA) single and ensemble methods and several SOTA meta-learners (adapted to our problem) in terms of selecting better forecasting models (i.e., 2X gain) for unseen tasks for univariate and multivariate testbeds.

****

PiCO: Contrastive Label Disambiguation for Partial Label Learning
Haobo Wang,Ruixuan Xiao,Yixuan Li,Lei Feng,Gang Niu,Gang Chen,Junbo Zhao

Partial label learning (PLL) is an important problem that allows each training example to be labeled with a coarse candidate set, which well suits many real-world data annotation scenarios with label ambiguity. Despite the promise, the performance of PLL often lags behind the supervised counterpart. In this work, we bridge the gap by addressing two key research challenges in PLL---representation learning and label disambiguation---in one coherent framework. Specifically, our proposed framework PiCO consists of a contrastive learning module along with a novel class prototype-based label disambiguation algorithm. PiCO produces closely aligned representations for examples from the same classes and facilitates label disambiguation. Theoretically, we show that these two components are mutually beneficial, and can be rigorously justified from an expectation-maximization (EM) algorithm perspective. Extensive experiments demonstrate that PiCO significantly outperforms the current state-of-the-art approaches in PLL and even achieves comparable results to fully supervised learning. Code and data available: https://github.com/hbzju/PiCO.

****

A Zest of LIME: Towards Architecture-Independent Model Distances
Hengrui Jia,Hongyu Chen,Jonas Guan,Ali Shahin Shamsabadi,Nicolas Papernot

Definitions of the distance between two machine learning models either characterize the similarity of the models' predictions or of their weights. While similarity of weights is attractive because it implies similarity of predictions in the limit, it suffers from being inapplicable to comparing models with different architectures. On the other hand, the similarity of predictions is broadly applicable but depends heavily on the choice of model inputs during comparison. In this paper, we instead propose to compute distance between black-box models by comparing their Local Interpretable Model-Agnostic Explanations (LIME). To compare two models, we take a reference dataset, and locally approximate the models on each reference point with linear models trained by LIME. We then compute the cosine distance between the concatenated weights of the linear models. This yields an approach that is both architecture-independent and possesses the benefits of comparing models in weight space. We empirically show that our method, which we call Zest, can be applied to two problems that require measurements of model similarity: detecting model stealing and machine unlearning.

****

Reconstructing Word Embeddings via Scattered k-Sub-Embedding
Soonyong Hwang,Byung-Ro Moon

The performance of modern neural language models relies heavily on the diversity of the vocabularies. Unfortunately, the language models tend to cover more vocabularies, the embedding parameters in the language models such as multilingual models used to occupy more than a half of their entire learning parameters. To solve this problem, we aim to devise a novel embedding structure to lighten the network without considerably performance degradation. To reconstruct N embedding vectors, we initialize k bundles of $M (\ll N)$ k-sub-embeddings to apply Cartesian product. Furthermore, we assign k-sub-embedding using the contextual relationship between tokens from pretrained language models. We adjust our k-sub-embedding structure to masked language models to evaluate proposed structure on downstream tasks. Our experimental results show that over 99.9$\%+$ compressed sub-embeddings for the language models performed comparably with the original embedding structure on GLUE and XNLI benchmarks.

****

Understanding Sharpness-Aware Minimization
Maksym Andriushchenko,Nicolas Flammarion

Sharpness-Aware Minimization (SAM) is a recent training method that relies on wo rst-case weight perturbations. SAM significantly improves generalization in various settings, however, existing justifications for its success do not seem conclusive. First, we analyze the implicit bias of SAM over diagonal linear networks, and prove that it always chooses a solution that enjoys better generalisation properties than standard gradient descent. We also provide a convergence proof of SAM for non-convex objectives when used with stochastic gradients and empirically discuss the convergence and generalization behavior of SAM for deep networks. Next, we discuss why SAM can be helpful in the noisy label setting where we first show that it can help to improve generalization even for linear classifiers. Then we discuss a gradient reweighting interpretation of SAM and show a further beneficial effect of combining SAM with a robust loss. Finally, we draw parallels between overfitting observed in learning with noisy labels and in adversarial training where SAM also improves generalization. This connection suggests that, more generally, techniques from the noisy label literature can be useful to improve robust generalization.

****

Path-specific Causal Fair Prediction via Auxiliary Graph Structure Learning

Liuyi Yao,Yaliang Li,Bolin Ding,Jingren Zhou,Jinduo Liu,Mengdi Huai,Jing Gao

Algorithm fairness has become a trending topic, and it has a great impact on social welfare. Among different fairness definitions, path-specific causal fairness is a widely adopted one with great potentials, as it distinguishes the fair and unfair effects that the sensitive attributes exert on algorithm predictions. Existing methods based on path-specific causal fairness either require graph structure as the prior knowledge or have high complexity in the calculation of path-specific effect. To tackle these challenges, we propose a novel casual graph based fair prediction framework, which integrates graph structure learning into fair prediction to ensure that unfair pathways are excluded in the causal graph. Furthermore, we generalize the proposed framework to the scenarios where sensitive attributes can be non-root nodes and affected by other variables, which is commonly observed in real-world applications but hardly addressed by existing works. We provide theoretical analysis on the generalization bound for the proposed fair prediction method, and conduct a series of experiments on real-world datasets to demonstrate that the proposed framework can provide better prediction performance and algorithm fairness trade-off.

****

Meta-Imitation Learning by Watching Video Demonstrations

Jiayi Li,Tao Lu,Xiaoge Cao,Yinghao Cai,Shuo Wang

Meta-Imitation Learning is a promising technique for the robot to learn a new task from observing one or a few human demonstrations. However, it usually requires a significant number of demonstrations both from humans and robots during the meta-training phase, which is a laborious and hard work for data collection, especially in recording the actions and specifying the correspondence between human and robot. In this work, we present an approach of meta-imitation learning by watching video demonstrations from humans. In comparison to prior works, our approach is able to translate human videos into practical robot demonstrations and train the meta-policy with adaptive loss based on the quality of the translated data. Our approach relies only on human videos and does not require robot demonstration, which facilitates data collection and is more in line with human imitation behavior. Experiments reveal that our method achieves the comparable performance to the baseline on fast learning a set of vision-based tasks through watching a single video demonstration.

****

VoiceFixer: Toward General Speech Restoration with Neural Vocoder

Haohe Liu,Qiuqiang Kong,Qiao Tian,Yan Zhao,DeLiang Wang,Chuanzeng Huang,Yuxuan Wang

Speech restoration aims to remove distortions in speech signals. Prior methods mainly focus on single-task speech restoration (SSR), such as speech denoising or speech declipping. However, SSR systems only focus on one task and do not address the general speech restoration problem. In addition, previous SSR systems show limited performance in some speech restoration tasks such as speech super-reso lution. To overcome those limitations, we propose a general speech restoration (GSR) task that attempts to remove multiple distortions ■simultaneously. Furthermore, we propose VoiceFixer, a generative framework to address the GSR task. VoiceFixer consists of an analysis stage and a synthesis stage to mimic the speech analysis and comprehension of the human auditory system. We employ a ResUNet to model the analysis stage and a neural vocoder to model the synthesis stage. We evaluate VoiceFixer with additive noise, room ■reverberation, low-resolution, and clipping distortions. Our baseline GSR model achieves a 0.499 higher mean opinion score (MOS) than the speech denoising SSR model. VoiceFixer further surpasses the GSR baseline model on the MOS score by 0.256. Moreover, we observe that VoiceFixer generalizes well to severely degraded real speech recordings, indicating its potential in restoring old movies and historical speeches.

****

DNN Quantization with Attention

Ghouthi BOUKLI HACENE,Lukas Mauch,Shubhankar Chowdhury,Stefan Uhlich,Fabien Cardinaux

Low-bit quantization of network weights and activations can drastically reduce the memory footprint, complexity, energy consumption and latency of Deep Neural Networks (DNNs). Many different quantization methods like min-max quantization, Statistics-Aware Weight Binning (SAWB) or Binary Weight Network (BWN) have been proposed in the past. However, they still cause a considerable accuracy drop, in particular when applied to complex learning tasks or lightweight DNN architectures. In this paper, we propose a novel training procedure that can be used to improve the performance of existing quantization methods. We call this procedure \textit{DNN Quantization with Attention} (DQA). It relaxes the training problem, using a learnable linear combination of high, medium and low-bit quantization at the beginning, while converging to a single low-bit quantization at the end of the training. We show empirically that this relaxation effectively smooths the loss function and therefore helps convergence. Moreover, we conduct experiments and show that our procedure improves the performance of many state-of-the-art quantization methods on various object recognition tasks. In particular, we apply DQA with min-max, SAWB and BWN to train $2$bit quantized DNNs on the CIFAR10, CIFAR100 and ImageNet ILSVRC 2012 datasets, achieving a very good accuracy comparing to other conterparts.

****

Language Model Pre-training on True Negatives

Zhuosheng Zhang,hai zhao,Masao Utiyama,Eiichiro Sumita

Discriminative pre-trained language models (PrLMs) learn to predict original texts from intentionally corrupted ones. Taking the former text as positive and the latter as negative samples, the discriminative PrLM can be trained effectively for contextualized representation. However, though the training of such a type of PrLMs highly relies on the quality of the automatically constructed samples, existing PrLMs simply treat all corrupted texts as equal negative without any examination, which actually lets the resulting model inevitably suffer from the false negative issue where training is carried out on wrong data and leads to less efficiency and less robustness in the resulting PrLMs.

Thus in this work, on the basis of defining the false negative issue in discriminative PrLMs that has been ignored for a long time, we design enhanced pre-training methods to counteract false negative predictions and encourage pre-training language models on true negatives, by correcting the harmful gradient updates subject to false negative predictions. Experimental results on GLUE and SQuAD benchmarks show that our counter-false-negative pre-training methods indeed bring about better performance together with stronger robustness.

****

Understanding Intrinsic Robustness Using Label Uncertainty

Xiao Zhang,David Evans

A fundamental question in adversarial machine learning is whether a robust classifier exists for a given task. A line of research has made some progress towards this goal by studying the concentration of measure, but we argue standard concentration fails to fully characterize the intrinsic robustness of a classificatio n problem since it ignores data labels which are essential to any classification task. Building on a novel definition of label uncertainty, we empirically demonstrate that error regions induced by state-of-the-art models tend to have much higher label uncertainty than randomly-selected subsets. This observation motivates us to adapt a concentration estimation algorithm to account for label uncertainty, resulting in more accurate intrinsic robustness measures for benchmark image classification problems.

****

Exploiting Minimum-Variance Policy Evaluation for Policy Optimization

Alberto Maria Metelli,Samuele Meta,Marcello Restelli

Off-policy methods are the basis of a large number of effective Policy Optimization (PO) algorithms. In this setting, Importance Sampling (IS) is typically employed as a what-if analysis tool, with the goal of estimating the performance of a target policy, given samples collected with a different behavioral policy. However, in Monte Carlo simulation, IS represents a variance minimization approach. In this field, a suitable behavioral distribution is employed for sampling, allowing diminishing the variance of the estimator below the one achievable when sampling from the target distribution. In this paper, we analyze IS in these two guises, showing the connections between the two objectives. We illustrate that variance minimization can be used as a performance improvement tool, with the advantage, compared with direct off-policy learning, of implicitly enforcing a trust region. We make use of these theoretical findings to build a PO algorithm, Policy Optimization via Optimal Policy Evaluation (PO2PE), that employs variance minimization as an inner loop. Finally, we present empirical evaluations on continuous RL benchmarks, with a particular focus on the robustness to small batch sizes.

****

Efficient Split-Mix Federated Learning for On-Demand and In-Situ Customization

Junyuan Hong,Haotao Wang,Zhangyang Wang,Jiayu Zhou

Federated learning (FL) provides a distributed learning framework for multiple participants to collaborate learning without sharing raw data. In many practical FL scenarios, participants have heterogeneous resources due to disparities in hardware and inference dynamics that require quickly loading models of different sizes and levels of robustness. The heterogeneity and dynamics together impose significant challenges to existing FL approaches and thus greatly limit FL's applicability. In this paper, we propose a novel Split-Mix FL strategy for heterogeneous participants that, once training is done, provides in-situ customization of model sizes and robustness. Specifically, we achieve customization by learning a set of base sub-networks of different sizes and robustness levels, which are later aggregated on-demand according to inference requirements. This split-mix strategy achieves customization with high efficiency in communication, storage, and inference. Extensive experiments demonstrate that our method provides better in-situ customization than the existing heterogeneous-architecture FL methods. Codes and pre-trained models are available: https://github.com/illidanlab/SplitMix.

****

TempoRL: Temporal Priors for Exploration in Off-Policy Reinforcement Learning

Marco Bagatella,Sammy Joe Christen,Otmar Hilliges

Effective exploration is a crucial challenge in deep reinforcement learning. Behavioral priors have been shown to tackle this problem successfully, at the expense of reduced generality and restricted transferability. We thus propose temporal priors as a non-Markovian generalization of behavioral priors for guiding exploration in reinforcement learning. Critically, we focus on state-independent temporal priors, which exploit the idea of temporal consistency and are generally applicable and capable of transferring across a wide range of tasks. We show how dynamically sampling actions from a probabilistic mixture of policy and temporal prior can accelerate off-policy reinforcement learning in unseen downstream tasks. We provide empirical evidence that our approach improves upon strong baselines in long-horizon continuous control tasks under sparse reward settings.

****

Gradient play in stochastic games: stationary points, convergence, and sample co mplexity

Runyu Zhang,Zhaolin Ren,Na Li

We study the performance of the gradient play algorithm for stochastic games (SGs), where each agent tries to maximize its own total discounted reward by making decisions independently based on current state information which is shared between agents. Policies are directly parameterized by the probability of choosing a certain action at a given state. We show that Nash equilibria (NEs) and first-order stationary policies are equivalent in this setting, and give a local convergence rate around strict NEs. Further, for a subclass of SGs called Markov potential games (which includes the cooperative setting with identical rewards among agents as an important special case), we design a sample-based reinforcement learning algorithm and give a non-asymptotic global convergence rate analysis for both exact gradient play and our sample-based learning algorithm. Our result shows that the number of iterations to reach an ϵ-NE scales linearly, instead of exponentially, with the number of agents. Local geometry and local stability are also considered, where we prove that strict NEs are local maxima of the total potential function and fully-mixed NEs are saddle points.

****

Are BERT Families Zero-Shot Learners? A Study on Their Potential and Limitations

Yue Wang,Lijun Wu,xiaobo liang,Juntao Li,Min Zhang

Starting from the resurgence of deep learning, language models (LMs) have never been so popular. Through simply increasing model scale and data size, large LMs pre-trained with self-supervision objectives demonstrate awe-inspiring results on both task performance and generalization. At the early stage, supervised fine-tuning is indispensable in adapting pre-trained language models (PLMs) to downstream tasks. Later on, the sustained growth of model capacity and data size, as well as newly presented pre-training techniques, make the PLMs perform well under the few-shot setting, especially in the recent paradigm of prompt-based learning. After witnessing the success of PLMs for few-shot tasks, we propose to further study the potential and limitations of PLMs for the zero-shot setting. We utilize 3 models from the most popular BERT family to launch the empirical study on 20 different datasets. We are surprised to find that a simple Multi-Null Prompting (without manually/automatically created prompts) strategy can yield very promising results on a few widely-used datasets, e.g., $86.59\%(\pm0.59)$ accuracy on the IMDB dataset, and $86.22\%(\pm2.71)$ accuracy on the Amazon dataset, which outperforms manually created prompts without engineering in achieving much better and stable performance with the accuracy of $74.06\%(\pm13.04)$, $75.54\%(\pm11.77)$ for comparison. However, we also observe some limitations of PLMs under the zero-shot setting, particularly for the language understanding tasks (e.g., GLUE).

****

Federated Robustness Propagation: Sharing Adversarial Robustness in Federated Learning

Junyuan Hong,Haotao Wang,Zhangyang Wang,Jiayu Zhou

Federated learning (FL) emerges as a popular distributed learning schema that learns a model from a set of participating users without requiring raw data to be shared. One major challenge of FL comes from heterogeneity in users, which may have distributionally different (or \emph{non-iid}) data and varying computation resources. Just like in centralized learning, FL users also desire model robustness against malicious attackers at test time. Whereas adversarial training (AT) provides a sound solution for centralized learning, extending its usage for FL users has imposed significant challenges, as many users may have very limited training data as well as tight computational budgets, to afford the data-hungry and costly AT. In this paper, we study a novel learning setting that propagates adversarial robustness from high-resource users that can afford AT, to those low-resource users that cannot afford it, during the FL process. We show that existing FL techniques cannot effectively propagate adversarial robustness among \emph{non-iid} users, and propose a simple yet effective propagation approach that transfers robustness through carefully designed batch-normalization statistics. We demonstrate the rationality and effectiveness of our method through extensive ex periments. Especially, the proposed method is shown to grant FL remarkable robustness even when only a small portion of users afford AT during learning.

****

Concentric Spherical GNN for 3D Representation Learning

James S Fox,Bo Zhao,Beatriz Gonzalez Del Rio,Siva Rajamanickam,Rampi Ramprasad,Le Song

Learning 3D representations of point clouds that generalize well to arbitrary orientations is a challenge of practical importance in problems ranging from computer vision to molecular modeling.

The proposed approach is based on a concentric spherical representation of 3D space, formed by nesting spatially-sampled spheres resulting from the highly regular icosahedral discretization.

We propose separate intra-sphere and inter-sphere convolutions over the resulting concentric spherical grid, which are combined into a convolutional framework for learning volumetric and rotationally equivariant representations over point clouds.

We demonstrate the effectiveness of our approach for 3D object classification, and towards resolving the electronic structure of atomistic systems.

****

Anti-Concentrated Confidence Bonuses For Scalable Exploration

Jordan T. Ash,Cyril Zhang,Surbhi Goel,Akshay Krishnamurthy,Sham M. Kakade

Intrinsic rewards play a central role in handling the exploration-exploitation tradeoff when designing sequential decision-making algorithms, in both foundational theory and state-of-the-art deep reinforcement learning. The LinUCB algorithm, a centerpiece of the stochastic linear bandits literature, prescribes an elliptical bonus which addresses the challenge of leveraging shared information in large action spaces. This bonus scheme cannot be directly transferred to high-dimensional exploration problems, however, due to the computational cost of maintaining the inverse covariance matrix of action features. We introduce anti-concentrated confidence bounds for efficiently approximating the elliptical bonus, using an ensemble of regressors trained to predict random noise from policy network-derived features. Using this approximation, we obtain stochastic linear bandit algorithms which obtain $\tilde O(d \sqrt{T})$ regret bounds for $\mathsf{poly}(d)$ fixed actions. We develop a practical variant that is competitive with contemporary intrinsic reward heuristics on Atari benchmarks.

****

Sqrt(d) Dimension Dependence of Langevin Monte Carlo

Ruilin Li,Hongyuan Zha,Molei Tao

This article considers the popular MCMC method of unadjusted Langevin Monte Carlo (LMC) and provides a non-asymptotic analysis of its sampling error in 2-Wasserstein distance. The proof is based on a refinement of mean-square analysis in Li et al. (2019), and this refined framework automates the analysis of a large class of sampling algorithms based on discretizations of contractive SDEs. Using this framework, we establish an $\tilde{O}(\sqrt{d}/\epsilon)$ mixing time bound for LMC, without warm start, under the common log-smooth and log-strongly-convex conditions, plus a growth condition on the 3rd-order derivative of the potential of target measures. This bound improves the best previously known $\tilde{O}(d/\epsilon)$ result and is optimal (in terms of order) in both dimension d and accuracy tolerance ϵ for target measures satisfying the aforementioned assumptions. Our theoretical analysis is further validated by numerical experiments.

****

Continual Learning with Filter Atom Swapping

Zichen Miao,Ze Wang,Wei Chen,Qiang Qiu

Continual learning has been widely studied in recent years to resolve the catastrophic forgetting of deep neural networks. In this paper, we first enforce a low-rank filter subspace by decomposing convolutional filters within each network layer over a small set of filter atoms. Then, we perform continual learning with filter atom swapping. In other words, we learn for each task a new filter subspace for each convolutional layer, i.e., hundreds of parameters as filter atoms, b ut keep subspace coefficients shared across tasks. By maintaining a small footprint memory of filter atoms, we can easily archive models for past tasks to avoid forgetting. The effectiveness of this simple scheme for continual learning is illustrated both empirically and theoretically. The proposed atom swapping framework further enables flexible and efficient model ensemble with members selected within a task or across tasks to improve the performance in different continual learning settings. Being validated on multiple benchmark datasets with different convolutional network structures, the proposed method outperforms the state-of-the-art methods in both accuracy and scalability.

****

Triangular Dropout: Variable Network Width without Retraining

Edward W Staley,Corban G Rivera,Neil Joshi

One of the most fundamental design choices in neural networks is layer width: it affects the capacity of what a network can learn and determines the complexity of the solution. This latter property is often exploited when introducing information bottlenecks, forcing a network to learn compressed representations. However, such an architecture decision is typically immutable once training begins; switching to a more compressed architecture requires retraining. In this paper we present a new layer design, called Triangular Dropout, which does not have this limitation. After training, the layer can be arbitrarily reduced in width to exchange performance for narrowness. We demonstrate the construction and potential use cases of such a mechanism in three areas. Firstly, we describe the formulation of Triangular Dropout in autoencoders, creating an MNIST autoencoder with selectable compression after training. Secondly, we add Triangular Dropout to VGG19 on ImageNet, creating a powerful network which, without retraining, can be significantly reduced in parameters with only small changes to classification accuracy. Lastly, we explore the application of Triangular Dropout to reinforcement learning (RL) policies on selected control problems, showing that it can be used to characterize the complexity of RL tasks, a critical measurement in multitask learning and lifelong-learning domains.

****

Relational Surrogate Loss Learning

Tao Huang,Zekang Li,Hua Lu,Yong Shan,Shusheng Yang,Yang Feng,Fei Wang,Shan You,Chang Xu

Evaluation metrics in machine learning are often hardly taken as loss functions, as they could be non-differentiable and non-decomposable, e.g., average precision and F1 score. This paper aims to address this problem by revisiting the surrogate loss learning, where a deep neural network is employed to approximate the evaluation metrics. Instead of pursuing an exact recovery of the evaluation metric through a deep neural network, we are reminded of the purpose of the existence of these evaluation metrics, which is to distinguish whether one model is better or worse than another. In this paper, we show that directly maintaining the relation of models between surrogate losses and metrics suffices, and propose a rank correlation-based optimization method to maximize this relation and learn surrogate losses. Compared to previous works, our method is much easier to optimize and enjoys significant efficiency and performance gains. Extensive experiments show that our method achieves improvements on various tasks including image classification and neural machine translation, and even outperforms state-of-the-art methods on human pose estimation and machine reading comprehension tasks. Code is available at: https://github.com/hunto/ReLoss.

****

Learning Rate Grafting: Transferability of Optimizer Tuning

Naman Agarwal,Rohan Anil,Elad Hazan,Tomer Koren,Cyril Zhang

In the empirical science of training large neural networks, the learning rate schedule is a notoriously challenging-to-tune hyperparameter, which can depend on all other properties (architecture, optimizer, batch size, dataset, regularization, ...) of the problem. In this work, we probe the entanglements between the optimizer and the learning rate schedule. We propose the technique of optimizer grafting, which allows for the transfer of the overall implicit step size schedule from a tuned optimizer to a new optimizer, preserving empirical performance. Th is provides a robust plug-and-play baseline for optimizer comparisons, leading to reductions to the computational cost of optimizer hyperparameter search. Using grafting, we discover a non-adaptive learning rate correction to SGD which allows it to train a BERT model to state-of-the-art performance. Besides providing a resource-saving tool for practitioners, the invariances discovered via grafting shed light on the successes and failure modes of optimizers in deep learning.

****

Structure-Aware Transformer Policy for Inhomogeneous Multi-Task Reinforcement Learning

Sunghoon Hong,Deunsol Yoon,Kee-Eung Kim

Modular Reinforcement Learning, where the agent is assumed to be morphologically structured as a graph, for example composed of limbs and joints, aims to learn a policy that is transferable to a structurally similar but different agent. Compared to traditional Multi-Task Reinforcement Learning, this promising approach allows us to cope with inhomogeneous tasks where the state and action space dimensions differ across tasks. Graph Neural Networks are a natural model for representing the pertinent policies, but a recent work has shown that their multi-hop message passing mechanism is not ideal for conveying important information to other modules and thus a transformer model without morphological information was proposed. In this work, we argue that the morphological information is still very useful and propose a transformer policy model that effectively encodes such information. Specifically, we encode the morphological information in terms of the traversal-based positional embedding and the graph-based relational embedding. We empirically show that the morphological information is crucial for modular reinforcement learning, substantially outperforming prior state-of-the-art methods on multi-task learning as well as transfer learning settings with different state and action space dimensions.

****

Toward Efficient Low-Precision Training: Data Format Optimization and Hysteresis Quantization

Sunwoo Lee,Jeongwoo Park,Dongsuk Jeon

As the complexity and size of deep neural networks continue to increase, low-precision training has been extensively studied in the last few years to reduce hardware overhead. Training performance is largely affected by the numeric formats representing different values in low-precision training, but finding an optimal format typically requires numerous training runs, which is a very time-consuming process. In this paper, we propose a method to efficiently find an optimal format for activations and errors without actual training. We employ this method to determine an 8-bit format suitable for training various models. In addition, we propose hysteresis quantization to suppress undesired fluctuation in quantized weights during training. This scheme enables deeply quantized training using 4-bit weights, exhibiting only 0.2% degradation for ResNet-18 trained on ImageNet.

****

Pyraformer: Low-Complexity Pyramidal Attention for Long-Range Time Series Modeling and Forecasting

Shizhan Liu,Hang Yu,Cong Liao,Jianguo Li,Weiyao Lin,Alex X. Liu,Schahram Dustdar

Accurate prediction of the future given the past based on time series data is of paramount importance, since it opens the door for decision making and risk management ahead of time. In practice, the challenge is to build a flexible but parsimonious model that can capture a wide range of temporal dependencies. In this paper, we propose Pyraformer by exploring the multiresolution representation of the time series. Specifically, we introduce the pyramidal attention module (PAM) in which the inter-scale tree structure summarizes features at different resolutions and the intra-scale neighboring connections model the temporal dependencies of different ranges. Under mild conditions, the maximum length of the signal traversing path in Pyraformer is a constant (i.e., $\mathcal O(1)$) with regard to the sequence length L, while its time and space complexity scale linearly with L. Extensive numerical results show that Pyraformer typically achieves the highest prediction accuracy in both single-step and long-range forecasting tasks with the least amount of time and memory consumption, especially when the sequen ce is long.

****

Knowledge Infused Decoding

Ruibo Liu,Guoqing Zheng,Shashank Gupta,Radhika Gaonkar,Chongyang Gao,Soroush Vosoughi,Milad Shokouhi,Ahmed Hassan Awadallah

Pre-trained language models (LMs) have been shown to memorize a substantial amount of knowledge from the pre-training corpora; however, they are still limited in recalling factually correct knowledge given a certain context. Hence. they tend to suffer from counterfactual or hallucinatory generation when used in knowledge-intensive natural language generation (NLG) tasks. Recent remedies to this problem focus on modifying either the pre-training or task fine-tuning objectives to incorporate knowledge, which normally require additional costly training or architecture modification of LMs for practical applications.

We present Knowledge Infused Decoding (KID)---a novel decoding algorithm for generative LMs, which dynamically infuses external knowledge into each step of the LM decoding. Specifically, we maintain a local knowledge memory based on the current context, interacting with a dynamically created external knowledge trie, and continuously update the local memory as a knowledge-aware constraint to guide decoding via reinforcement learning. On six diverse knowledge-intensive NLG tasks, task-agnostic LMs (e.g., GPT-2 and BART) armed with KID outperform many task-optimized state-of-the-art models, and show particularly strong performance in few-shot scenarios over seven related knowledge-infusion techniques. Human evaluation confirms KID's ability to generate more relevant and factual language for the input context when compared with multiple baselines. Finally, KID also alleviates exposure bias and provides stable generation quality when generating longer sequences.

****

Parallel Training of GRU Networks with a Multi-Grid Solver for Long Sequences

Euhyun Moon,Eric C Cyr

Parallelizing Gated Recurrent Unit (GRU) is a challenging task, as the training procedure of GRU is inherently sequential. Prior efforts to parallelize GRU have largely focused on conventional parallelization strategies such as data-parallel and model-parallel training algorithms. However, when the given sequences are very long, existing approaches are still inevitably performance limited in terms of both training time and model accuracy. In this paper, we present a novel parallel training scheme (called parallel-in-time) for GRU based on a multigrid reduction in time (MGRIT) solver. MGRIT partitions a sequence into multiple shorter sub-sequences and trains the sub-sequences on different processors in parallel. The key to achieving speedup is a hierarchical correction of the hidden state to accelerate end-to-end communication in both the forward and backward propagation phases of gradient descent. Experimental results on the HMDB51 dataset, where each video is an image sequence, demonstrate that a new parallel training scheme of GRU achieves up to $6.5 \times$ speedup over a serial approach. As efficiency of our new parallelization strategy is associated with the sequence length, our parallel GRU algorithm achieves significant performance improvement as the length of sequence increases. Further, the proposed approach can be applied simultaneously with batch and other forms of model parallelism.

****

Stability and Generalisation in Batch Reinforcement Learning

Matthew J. A. Smith,Shimon Whiteson

Overfitting has been recently acknowledged as a key limiting factor in the capabilities of reinforcement learning algorithms, despite little theoretical characterisation. We provide a theoretical examination of overfitting in the context of batch reinforcement learning, through the fundamental relationship between algorithmic stability (Bousquet & Elisseeff, 2002)–which characterises the effect of a change at a single data point–and the generalisation gap–which quantifies overfitting. Examining a popular fitted policy evaluation method with linear value function approximation, we characterise the dynamics of overfitting in the RL context. We provide finite sample, finite time, polynomial bounds on the generalis ation gap in RL. In addition, our approach applies to a class of algorithms which only partially fit to temporal difference errors, as is common in deep RL, rather than perfectly optimising at each step. As such, our results characterise an unexplored bias-variance trade-off in the frequency of target network updates. To do so, our work extends the stochastic gradient-based approach of Hardt et al. (2016) to the iterative methods more common in RL. We find that under regimes where learning requires few iterations, the expected temporal difference error over the dataset is representative of the true performance on the MDP, indicating that, as is the case in supervised learning, good generalisation in RL can be ensured through the use of algorithms that learn quickly.

****

Analyzing the Effects of Classifier Lipschitzness on Explainers

Zulqarnain Khan,Aria Masoomi,Davin Hill,Jennifer Dy

Machine learning methods are getting increasingly better at making predictions, but at the same time they are also becoming more complicated and less transparent. As a result, explanation methods are often relied on to provide interpretability to these complicated and often black-box prediction models. As crucial diagnostics tools, it is important that these explainer methods themselves are reliable. In this paper we focus on one particular aspect of reliability, namely that an explainer should give similar explanations for similar data inputs. We formalize this notion by introducing and defining explainer astuteness, analogous to astuteness of classifiers. Our formalism is inspired by the concept of probabilistic Lipschitzness, which captures the probability of local smoothness of a function. For a variety of explainers (e.g., SHAP, RISE, CXPlain, PredDiff), we provide lower bound guarantees on the astuteness of these explainers given the Lipschitzness of the prediction function. These theoretical results imply that locally smooth prediction functions lend themselves to locally robust explanations. We evaluate these results empirically on simulated as well as real datasets.

****

Contrastive Embeddings for Neural Architectures

Daniel Hesslow,Iacopo Poli

The performance of algorithms for neural architecture search strongly depends on the parametrization of the search space. We use contrastive learning to identify networks across different initializations based on their data Jacobians and their number of parameters, and automatically produce the first architecture embeddings independent from the parametrization of the search space. Using our contrastive embeddings, we show that traditional black-box optimization algorithms, without modification, can reach state-of-the-art performance in Neural Architecture Search. As our method provides a unified embedding space, we successfully perform transfer learning between search spaces. Finally, we show the evolution of embeddings during training, motivating future studies into using embeddings at different training stages to gain a deeper understanding of the networks in a search space.

****

QUERY EFFICIENT DECISION BASED SPARSE ATTACKS AGAINST BLACK-BOX DEEP LEARNING MODELS

Viet Vo,Ehsan M Abbasnejad,Damith Ranasinghe

Despite our best efforts, deep learning models remain highly vulnerable to even tiny adversarial perturbations applied to the inputs. The ability to extract information from solely the output of a machine learning model to craft adversarial perturbations to black-box models is a practical threat against real-world systems, such as Machine Learning as a Service (MLaaS), particularly $sparse~attacks$. The realization of sparse attacks in black-box settings demonstrates that machine learning models are more vulnerable than we believe. Because, these attacks aim to $minimize~the~number~of~perturbed~pixels$—measured by l_0 norm—required to mislead a model by $solely$ observing the decision ($the~predicted~label$) returned to a model query; the so-called $decision-based~setting$. But, such an attack leads to an NP-hard optimization problem. We develop an evolution-based algorithm—$SparseEvo$—for the problem and evaluate it against both convolutional deep neural networks and $vision~transformers$. Notably, vision transformers are yet to be investigated under a decision-based attack setting. SparseEvo requires significantly fewer queries than the state-of-the-art sparse attack $Pointwise$ for both untargeted and targeted attacks. The attack algorithm, although conceptually simple, is competitive with only a limited query budget against the state-of-the-art gradient-based $white-box$ attacks in standard computer vision tasks such as $ImageNet$. Importantly, the query efficient SparseEvo, along with decision-based attacks, in general, raises new questions regarding the safety of deployed systems and poses new directions to study and understand the robustness of machine learning models.

****

Learning Invariant Reward Functions through Trajectory Interventions

Ivan Ovinnikov,Eugene Bykovets,Joachim M. Buhmann

Inverse reinforcement learning methods aim to retrieve the reward function of a Markov decision process based on a dataset of expert demonstrations. The commonplace scarcity of such demonstrations potentially leads to the absorption of spurious correlations in the data by the learning model, which as a result, exhibits behavioural overfitting to the expert dataset when trained on the obtained reward function. We study the generalization properties of the maximum entropy method for solving the inverse reinforcement learning problem for both exact and approximate formulations and demonstrate that by applying an instantiation of the invariant risk minimization principle, we can recover reward functions which induce better performing policies across domains in the transfer setting.

****

On the Convergence of Nonconvex Continual Learning with Adaptive Learning Rate

Sungyeob Han,Yeongmo Kim,Jungwoo Lee

One of the objectives of continual learning is to prevent catastrophic forgetting in learning multiple tasks sequentially.
The memory based continual learning stores a small subset of the data for previous tasks and applies various methods such as quadratic programming and sample selection.
Some memory-based approaches are formulated as a constrained optimization problem and rephrase constraints on the objective for memory as the inequalities on gradients.
However, there have been little theoretical results on the convergence of continual learning.
In this paper, we propose a theoretical convergence analysis of memory-based continual learning with stochastic gradient descent.
The proposed method called nonconvex continual learning (NCCL) adapts the learning rates of both previous and current tasks with the gradients.
The proposed method can achieve the same convergence rate as the SGD method for a single task when the catastrophic forgetting term which we define in the paper is suppressed at each iteration.
It is also shown that memory-based approaches inherently overfit to memory, which degrades the performance on previously learned tasks. Experiments show that the proposed algorithm improves the performance of continual learning over existing methods for several image classification tasks.

****

Almost Tight L0-norm Certified Robustness of Top-k Predictions against Adversarial Perturbations

Jinyuan Jia,Binghui Wang,Xiaoyu Cao,Hongbin Liu,Neil Zhenqiang Gong

Top-k predictions are used in many real-world applications such as machine learning as a service, recommender systems, and web searches. ℓ_0-norm adversarial perturbation characterizes an attack that arbitrarily modifies some features of an input such that a classifier makes an incorrect prediction for the perturbed input. ℓ_0-norm adversarial perturbation is easy to interpret and can be implemented in the physical world. Therefore, certifying robustness of top-k predictions against ℓ_0-norm adversarial perturbation is important. However, existing studies either focused on certifying ℓ_0-norm robustness of top-1 predictions or ℓ_2-norm robustness of top-k predictions. In this work, we aim to bridge the gap. Our approach is based on randomized smoothing, which builds a provably robust classifier from an arbitrary classifier via randomizing an input. Our major theoretical contribution is an almost tight ℓ_0-norm certified robustness guarantee for top-k predictions. We empirically evaluate our method on CIFAR10 and ImageNet. For instance, our method can build a classifier that achieves a certified top-3 accuracy of 69.2\% on ImageNet when an attacker can arbitrarily perturb 5 pixels of a testing image.

****

Direct Molecular Conformation Generation

Jinhua Zhu,Yingce Xia,Chang Liu,Lijun Wu,Shufang Xie,Wengang Zhou,Tao Qin,Houqiang Li,Tie-Yan Liu

Molecular conformation generation, which is to generate 3 dimensional coordinates of all the atoms in a molecule, is an important task for bioinformatics and pharmacology. Most existing machine learning based methods first predict interatomic distances and then generate conformations based on them. This two-stage approach has a potential limitation that the predicted distances may conflict with each other, e.g., violating the triangle inequality. In this work, we propose a method that directly outputs the coordinates of atoms, so that there is no violation of constraints. The conformation generator of our method stacks multiple blocks, and each block outputs a conformation which is then refined by the following block. We adopt the variational auto-encoder (VAE) framework and use a latent variable to generate diverse conformations. To handle the roto-translation equivariance, we adopt a loss that is invariant to rotation and translation of molecule coordinates, by computing the minimal achievable distance after any rotation and translation. Our method outperforms strong baselines on four public datasets, which shows the effectiveness of our method and the great potential of the direct approach. The code is released at \url{https://github.com/DirectMolecularConfGen/DMCG}.

****

Generalization in Deep RL for TSP Problems via Equivariance and Local Search

Wenbin Ouyang,Yisen Wang,Paul Weng,Shaochen Han

Deep reinforcement learning (RL) has proved to be a competitive heuristic for solving small-sized instances of traveling salesman problems (TSP), but its performance on larger-sized instances is insufficient. Since training on large instances is impractical, we design a novel deep RL approach with a focus on generalizability. Our proposition consisting of a simple deep learning architecture that learns with novel RL training techniques exploits two main ideas. First, we exploit equivariance to facilitate training. Second, we interleave efficient local search heuristics with the usual RL training to smooth the value landscape. In order to validate the whole approach, we empirically evaluate our proposition on random and realistic TSP problems against relevant state-of-the-art deep RL methods. Moreover, we present an ablation study to understand the contribution of each of its components.

****

ScheduleNet: Learn to solve multi-agent scheduling problems with reinforcement learning

Junyoung Park,Sanzhar Bakhtiyarov,Jinkyoo Park

We propose ScheduleNet, an RL-based decentralized constructive scheduler for coordinating multi-agent to finish tasks with minimum completion time. We formulate multi-agent scheduling problems (mSPs) as an event-based Markov decision process (MDP) with an episodic reward (e.g., makespan) and derive a decentralized decision-making policy using reinforcement learning. The decision making procedure of ScheduleNet includes: (1) representing the state of a scheduling problem with the agent-task graph, (2) extracting node embeddings for agent and tasks nodes by employing the type-aware graph attention (TGA), and (3) computing the assignment probability with the computed node embeddings. We validate the effectiveness of ScheduleNet on two types of mSPs: multiple traveling salesmen problem (mTSP) and job-shop scheduling problem (JSP). We empirically show that ScheduleNet can outperform other heuristic approaches and existing deep RL approaches, particularly validating its exceptional effectiveness in solving large and practical problems. Furthermore, we have demonstrated that ScheduleNet can effectively solve online vehicle routing problems where the new target customer appears dynamically during the course of scheduling.

****

Proving the Lottery Ticket Hypothesis for Convolutional Neural Networks
Arthur da Cunha,Emanuele Natale,Laurent Viennot
The lottery ticket hypothesis states that a randomly-initialized neural network contains a small subnetwork which, when trained in isolation, can compete with the performance of the original network. Recent theoretical works proved an even stronger version: every sufficiently overparameterized (dense) neural network contains a subnetwork that, even without training, achieves accuracy comparable to that of the trained large network. These works left as an open problem to extend the result to convolutional neural networks (CNNs).
In this work we provide such generalization by showing that, with high probability, it is possible to approximate any CNN by pruning a random CNN whose size is larger by a logarithmic factor.

****

Neurally boosted supervised spectral clustering
Ali Parviz,Ioannis Koutis
Network embedding methods compute geometric representations of graphs that render various prediction problems amenable to machine learning techniques. Spectral network embeddings are based on the computation of eigenvectors of a normalized graph Laplacian. When coupled with standard classifiers, spectral embeddings yield strong baseline performance in node classification tasks. Remarkably, it has been recently shown that these `base' classifications followed by a simple `Correction and Smooth' procedure reach state-of-the-art performance on widely used benchmarks. All these recent works employ classifiers that are agnostic to the nature of the underlying embedding. We present simple neural models that leverage fundamental geometric properties of spectral embeddings and obtains significantly improved classification accuracy over commonly used standard classifiers. Our results are based on a specific variant of spectral clustering that is not well-known, but it is presently the only variant known to have analyzable theoretical properties. We provide a \texttt{PyTorch} implementation of our classifier along with code for the fast computation of spectral embeddings.

****

Label Refining: a semi-supervised method to extract voice characteristics without ground truth
Mathias Quillot,Richard Dufour,Jean-françois Bonastre
A characteristic is a distinctive trait shared by a group of observations which may be used to identify them. In the context of voice casting for audiovisual productions, characteristic extraction has an important role since it can help explaining the decisions of a voice recommendation system, or give modalities to the user with the aim to express a voice search request. Unfortunately, the lack of standard taxonomy to describe comedian voices prevents the implementation of an annotation protocol. To address this problem, we propose a new semi-supervised learning method entitled Label Refining that consists in extracting refined labels (e.g. vocal characteristics) from known initial labels (e.g. character played in a recording). Our proposed method first suggests using a representation extractor based on the initial labels, then computing refined labels using a clustering algorithm to finally train a refined representation extractor. The method is validated by applying Label Refining on recordings from the video game MassEffect 3. Experiments show that, using a subsidiary corpus, it is possible to bring out interesting voice characteristics without any a priori knowledge.

****

Discovering Nonlinear PDEs from Scarce Data with Physics-encoded Learning
Chengping Rao,Pu Ren,Yang Liu,Hao Sun

There have been growing interests in leveraging experimental measurements to discover the underlying partial differential equations (PDEs) that govern complex physical phenomena. Although past research attempts have achieved great success in data-driven PDE discovery, the robustness of the existing methods cannot be guaranteed when dealing with low-quality measurement data. To overcome this challenge, we propose a novel physics-encoded discrete learning framework for discovering spatiotemporal PDEs from scarce and noisy data. The general idea is to (1) firstly introduce a novel deep convolutional-recurrent networks, which can encode prior physics knowledge (e.g., known terms, assumed PDE structure, initial/boundary conditions, etc.) while remaining flexible on representation capability, to accurately reconstruct high-fidelity data, and (2) then perform sparse regression with the reconstructed data to identify the analytical form of the governing PDEs. We validate our proposed framework on three high-dimensional PDE systems. The effectiveness and superiority of the proposed method over baselines are demonstrated.
****

Boosting Search Engines with Interactive Agents

Leonard Adolphs,Benjamin Börschinger,Christian Buck,Michelle Chen Huebscher,Massimiliano Ciaramita,Lasse Espeholt,Thomas Hofmann,Yannic Kilcher,Sascha Rothe,Pier Giuseppe Sessa,Lierni Sestorain

This paper presents first successful steps in designing agents that learn meta-strategies for iterative query refinement.
Our approach uses machine reading to guide the selection of refinement terms from aggregated search results.

Agents are then empowered with simple but effective search operators to exert fine-grained and transparent control over queries and search results.

We develop a novel way of generating synthetic search sessions, which leverages the power of transformer-based language models through (self-)supervised learning. We also present a reinforcement learning agent with dynamically constrained actions that learns interactive search strategies from scratch.

We obtain retrieval and answer quality performance comparable to recent neural methods using a traditional term-based BM25 ranking function. We provide an in-depth analysis of the search policies.
****

Examining Scaling and Transfer of Language Model Architectures for Machine Translation

Biao Zhang,Behrooz Ghorbani,Ankur Bapna,Yong Cheng,Xavier Garcia,Jonathan Shen,Orhan Firat

Natural language understanding and generation models follow one of the two dominant architectural paradigms: language models (LMs) that process concatenated sequences in a single stack of layers, and encoder-decoder models (EncDec) that utilize separate layer stacks for input and output processing. In machine translation, EncDec has long been the favoured approach, but with few studies investigating the performance of LMs. In this work, we thoroughly examine the role of several architectural design choices on the performance of LMs on bilingual, (massively) multilingual and zero-shot translation tasks, under systematic variations of data conditions and model sizes. Our results show that: (i) Different LMs have different scaling properties, where architectural differences often have a significant impact on model performance at small scales, but the performance gap narrows as the number of parameters increases, (ii) Several design choices, including causal masking and language-modeling objectives for the source sequence, have detrimental effects on translation quality, and (iii) When paired with full-visible masking for source sequences, LMs could perform on par with EncDec on supervised bilingual and multilingual translation tasks, but improve greatly on zero-shot directions by facilitating the reduction of off-target translations.
****

Rethinking Goal-Conditioned Supervised Learning and Its Connection to Offline RL

Rui Yang,Yiming Lu,Wenzhe Li,Hao Sun,Meng Fang,Yali Du,Xiu Li,Lei Han,Chongjie Zhang

Solving goal-conditioned tasks with sparse rewards using self-supervised learning is promising because of its simplicity and stability over current reinforcement learning (RL) algorithms. A recent work, called Goal-Conditioned Supervised Learning (GCSL), provides a new learning framework by iteratively relabeling and imitating self-generated experiences. In this paper, we revisit the theoretical property of GCSL --- optimizing a lower bound of the goal reaching objective, and extend GCSL as a novel offline goal-conditioned RL algorithm. The proposed method is named Weighted GCSL (WGCSL), in which we introduce an advanced compound weight consisting of three parts (1) discounted weight for goal relabeling, (2) goal-conditioned exponential advantage weight, and (3) best-advantage weight. Theoretically, WGCSL is proved to optimize an equivalent lower bound of the goal-conditioned RL objective and generates monotonically improved policies via an iterated scheme. The monotonic property holds for any behavior policies, and therefore WGCSL can be applied to both online and offline settings. To evaluate algorithms in the offline goal-conditioned RL setting, we provide a benchmark including a range of point and simulated robot domains. Experiments in the introduced benchmark demonstrate that WGCSL can consistently outperform GCSL and existing state-of-the-art offline methods in the fully offline goal-conditioned setting.

****

Topologically Regularized Data Embeddings

Robin Vandaele,Bo Kang,Jefrey Lijffijt,Tijl De Bie,Yvan Saeys

Unsupervised feature learning often finds low-dimensional embeddings that capture the structure of complex data. For tasks for which prior expert topological knowledge is available, incorporating this into the learned representation may lead to higher quality embeddings. For example, this may help one to embed the data into a given number of clusters, or to accommodate for noise that prevents one from deriving the distribution of the data over the model directly, which can then be learned more effectively. However, a general tool for integrating different prior topological knowledge into embeddings is lacking. Although differentiable topology layers have been recently developed that can (re)shape embeddings into prespecified topological models, they have two important limitations for representation learning, which we address in this paper. First, the currently suggested topological losses fail to represent simple models such as clusters and flares in a natural manner. Second, these losses neglect all original structural (such as neighborhood) information in the data that is useful for learning. We overcome these limitations by introducing a new set of topological losses, and proposing their usage as a way for topologically regularizing data embeddings to naturally represent a prespecified model. We include thorough experiments on synthetic and real data that highlight the usefulness and versatility of this approach, with applications ranging from modeling high-dimensional single-cell data, to graph embedding.

****

PF-GNN: Differentiable particle filtering based approximation of universal graph representations

Mohammed Haroon Dupty,Yanfei Dong,Wee Sun Lee

Message passing Graph Neural Networks (GNNs) are known to be limited in expressive power by the 1-WL color-refinement test for graph isomorphism. Other more expressive models either are computationally expensive or need preprocessing to extract structural features from the graph. In this work, we propose to make GNNs universal by guiding the learning process with exact isomorphism solver techniques which operate on the paradigm of $\textit{Individualization and refinement}$ (IR), a method to artificially introduce asymmetry and further refine the coloring when 1-WL stops. Isomorphism solvers generate a search-tree of colorings whose leaves uniquely identify the graph. However, the tree grows exponentially large and needs hand-crafted pruning techniques which are not desirable from a learning perspective. We take a probabilistic view and approximate the search tree of colorings (i.e. embeddings) by sampling multiple paths from root to leaves of the search-tree. To learn more discriminative representations, we guide the sampl ing process with $\textit{particle filter}$ updates, a principled approach for sequential state estimation. Our algorithm is end-to-end differentiable, can be applied with any GNN as backbone and learns richer graph representations with only linear increase in runtime. Experimental evaluation shows that our approach consistently outperforms leading GNN models on both synthetic benchmarks for isomorphism detection as well as real-world datasets.

****

Disentangling Sources of Risk for Distributional Multi-Agent Reinforcement Learning

Kyunghwan Son,Junsu Kim,Yung Yi,Jinwoo Shin

In cooperative multi-agent reinforcement learning, state transitions, rewards, and actions can all induce randomness (or uncertainty) in the observed long-term returns. These randomnesses are reflected from two risk sources: (a) agent-wise risk (i.e., how cooperative our teammates act for a given agent) and (b) environment-wise risk (i.e., transition stochasticity). Although these two sources are both important factors for learning robust policies of agents, prior works do not separate them or deal with only a single risk source, which could lead to suboptimal equilibria. In this paper, we propose Disentangled RIsk-sensitive Multi-Agent reinforcement learning (DRIMA), a novel framework being capable of disentangling risk sources. Our main idea is to separate risk-level leverages (i.e., quantiles) in both centralized training and decentralized execution with a hierarchical quantile structure and quantile regression. Our experiments demonstrate that DRIMA significantly outperforms prior-arts across various scenarios in StarCraft Multi-agent Challenge. Notably, DRIMA shows robust performance regardless of reward shaping, exploration schedule, where prior methods learn only a suboptimal policy.

****

Nonlinear ICA Using Volume-Preserving Transformations

Xiaojiang Yang,Yi Wang,Jiacheng Sun,Xing Zhang,Shifeng Zhang,Zhenguo Li,Junchi Yan

Nonlinear ICA is a fundamental problem in machine learning, aiming to identify the underlying independent components (sources) from data which is assumed to be a nonlinear function (mixing function) of these sources. Recent works prove that if the sources have some particular structures (e.g. temporal structure), they are theoretically identifiable even if the mixing function is arbitrary. However, in many cases such restrictions on the sources are difficult to satisfy or even verify, hence it inhibits the applicability of the proposed methods. Different from these works, we propose a general framework for nonlinear ICA, in which the mixing function is assumed to be a volume-preserving transformation, and meanwhile the conditions on the sources can be much looser. We provide an insightful proof of the identifiability of the proposed framework. We implement the framework by volume-preserving Flow-based models, and verify our theory by experiments on artificial data and synthesized images. Moreover, results on real-world images indicate that our framework can disentangle interpretable features.

****

Online Ad Hoc Teamwork under Partial Observability

Pengjie Gu,Mengchen Zhao,Jianye Hao,Bo An

Autonomous agents often need to work together as a team to accomplish complex cooperative tasks. Due to privacy and other realistic constraints, agents might need to collaborate with previously unknown teammates on the fly. This problem is known as ad hoc teamwork, which remains a core research challenge. Prior works usually rely heavily on strong assumptions like full observability, fixed and predefined teammates' types. This paper relaxes these assumptions with a novel reinforcement learning framework called ODITS, which allows the autonomous agent to adapt to arbitrary teammates in an online fashion. Instead of limiting teammates into a finite set of predefined types, ODITS automatically learns latent variables of teammates' behaviors to infer how to cooperate with new teammates effectively. To overcome partial observability, we introduce an information-based regularizer to derive proxy representations of the learned variables from local observations. Extensive experimental results show that ODITS significantly outperform

s various baselines in widely used ad hoc teamwork tasks.
****

Normalized Attention Without Probability Cage
Oliver Paul Richter,Roger Wattenhofer

Despite the popularity of attention based architectures like Transformers, the geometrical implications of softmax-attention remain largely unexplored. In this work we highlight the limitations of constraining attention weights to the probability simplex and the resulting convex hull of value vectors. We show that Transformers are biased towards local information at initialization and sensitive to hyperparameters, contrast attention to max- and sum-pooling and show the performance implications of different architectures with respect to biases in the data. Finally, we propose to replace the softmax in self-attention with normalization, resulting in a generally applicable architecture that is robust to hyperparameters and biases in the data. We support our insights with empirical results from more than 30,000 trained models. Implementations are in the supplementary material.
****

Continual Learning with Recursive Gradient Optimization
Hao Liu,Huaping Liu

Learning multiple tasks sequentially without forgetting previous knowledge, called Continual Learning(CL), remains a long-standing challenge for neural networks. Most existing methods rely on additional network capacity or data replay. In contrast, we introduce a novel approach which we refer to as Recursive Gradient Optimization(RGO). RGO is composed of an iteratively updated optimizer that modifies the gradient to minimize forgetting without data replay and a virtual Feature Encoding Layer(FEL) that represents different long-term structures with only task descriptors. Experiments demonstrate that RGO has significantly better performance on popular continual classification benchmarks when compared to the baselines and achieves new state-of-the-art performance on 20-split-CIFAR100(82.22%) and 20-split-miniImageNet(72.63%). With higher average accuracy than Single-Task Learning(STL), this method is flexible and reliable to provide continual learning capabilities for learning models that rely on gradient descent.
****

Continual Normalization: Rethinking Batch Normalization for Online Continual Learning
Quang Pham,Chenghao Liu,Steven HOI

Existing continual learning methods use Batch Normalization (BN) to facilitate training and improve generalization across tasks. However, the non-i.i.d and non-stationary nature of continual learning data, especially in the online setting, amplify the discrepancy between training and testing in BN and hinder the performance of older tasks. In this work, we study the cross-task normalization effect of BN in online continual learning where BN normalizes the testing data using moments biased towards the current task, resulting in higher catastrophic forgetting. This limitation motivates us to propose a simple yet effective method that we call Continual Normalization (CN) to facilitate training similar to BN while mitigating its negative effect. Extensive experiments on different continual learning algorithms and online scenarios show that CN is a direct replacement for BN and can provide substantial performance improvements. Our implementation will be made publicly available upon acceptance.
****

Equivariant Graph Mechanics Networks with Constraints
Wenbing Huang,Jiaqi Han,Yu Rong,Tingyang Xu,Fuchun Sun,Junzhou Huang

Learning to reason about relations and dynamics over multiple interacting objects is a challenging topic in machine learning. The challenges mainly stem from that the interacting systems are exponentially-compositional, symmetrical, and commonly geometrically-constrained.

Current methods, particularly the ones based on equivariant Graph Neural Networks (GNNs), have targeted on the first two challenges but remain immature for constrained systems.

In this paper, we propose Graph Mechanics Network (GMN) which is combinatorially efficient, equivariant and constraint-aware. The core of GMN is that it represents, by generalized coordinates, the forward kinematics information (positions and velocities) of a structural object. In this manner, the geometrical constraints are implicitly and naturally encoded in the forward kinematics. Moreover, to allow equivariant message passing in GMN, we have developed a general form of orthogonality-equivariant functions, given that the dynamics of constrained systems are more complicated than the unconstrained counterparts. Theoretically, the proposed equivariant formulation is proved to be universally expressive under certain conditions. Extensive experiments support the advantages of GMN compared to the state-of-the-art GNNs in terms of prediction accuracy, constraint satisfaction and data efficiency on the simulated systems consisting of particles, sticks and hinges, as well as two real-world datasets for molecular dynamics prediction and human motion capture.

****

Towards Continual Knowledge Learning of Language Models

Joel Jang,Seonghyeon Ye,Sohee Yang,Joongbo Shin,Janghoon Han,Gyeonghun KIM,Stanley Jungkyu Choi,Minjoon Seo

Large Language Models (LMs) are known to encode world knowledge in their parameters as they pretrain on a vast amount of web corpus, which is often utilized for performing knowledge-dependent downstream tasks such as question answering, fact-checking, and open dialogue. In real-world scenarios, the world knowledge stored in the LMs can quickly become outdated as the world changes, but it is non-trivial to avoid catastrophic forgetting and reliably acquire new knowledge while preserving invariant knowledge. To push the community towards better maintenance of ever-changing LMs, we formulate a new continual learning (CL) problem called Continual Knowledge Learning (CKL). We construct a new benchmark and metric to quantify the retention of time-invariant world knowledge, the update of outdated knowledge, and the acquisition of new knowledge. We adopt applicable recent methods from literature to create several strong baselines. Through extensive experiments, we find that CKL exhibits unique challenges that are not addressed in previous CL setups, where parameter expansion is necessary to reliably retain and learn knowledge simultaneously. By highlighting the critical causes of knowledge forgetting, we show that CKL is a challenging and important problem that helps us better understand and train ever-changing LMs.

****

Learning Pseudometric-based Action Representations for Offline Reinforcement Learning

Pengjie Gu,Mengchen Zhao,Chen Chen,Dong Li,Jianye Hao,Bo An

Offline reinforcement learning is a promising approach for practical applications since it does not require interactions with real-world environments. However, existing offline RL methods only work well in environments with continuous or small discrete action spaces. In environments with large and discrete action spaces, such as recommender systems and dialogue systems, the performance of existing methods decreases drastically because they suffer from inaccurate value estimation for a large proportion of o.o.d. actions. While recent works have demonstrated that online RL benefits from incorporating semantic information in action representations, unfortunately, they fail to learn reasonable relative distances between action representations, which is key to offline RL to reduce the influence of out-of-distribution (o.o.d.) actions. This paper proposes an action representation learning framework for offline RL based on a pseudometric, which measures both the behavioral relation and the data-distributional relation between actions. We provide theoretical analysis on the continuity and the bounds of the expected Q-values using the learned action representations. Experimental results show that our methods significantly improve the performance of two typical offline RL methods in environments with large and discrete action spaces.

****

NODE-GAM: Neural Generalized Additive Model for Interpretable Deep Learning

Chun-Hao Chang,Rich Caruana,Anna Goldenberg

Deployment of machine learning models in real high-risk settings (e.g. healthcare) often depends not only on the model's accuracy but also on its fairness, robu stness, and interpretability. Generalized Additive Models (GAMs) are a class of interpretable models with a long history of use in these high-risk domains, but they lack desirable features of deep learning such as differentiability and scalability. In this work, we propose a neural GAM (NODE-GAM) and neural GA^{2}M (NODE-GA^{2}M) that scale well and perform better than other GAMs on large datasets, while remaining interpretable compared to other ensemble and deep learning models. We demonstrate that our models find interesting patterns in the data. Lastly, we show that we are able to improve model accuracy via self-supervised pre-training, an improvement that is not possible for non-differentiable GAMs.

****

Localized Randomized Smoothing for Collective Robustness Certification

Jan Schuchardt,Tom Wollschläger,Aleksandar Bojchevski,Stephan Günnemann

Models for image segmentation, node classification and many other tasks map a single input to multiple labels. By perturbing this single shared input (e.g. the image) an adversary can manipulate several predictions (e.g. misclassify several pixels). A recent collective robustness certificate provides strong guarantees on the number of predictions that are simultaneously robust. This method is however limited to strictily models, where each prediction is associated with a small receptive field. We propose a more general collective certificate for the larger class of softly local models, where each output is dependent on the entire input but assigns different levels of importance to different input regions (e.g. based on their proximity in the image). The certificate is based on our novel localized randomized smoothing approach, where the random perturbation strength for different input regions is proportional to their importance for the outputs. The resulting locally smoothed model yields strong collective guarantees while maintaining high prediction quality on both image segmentation and node classification tasks.

****

Divide and Explore: Multi-Agent Separate Exploration with Shared Intrinsic Motivations

Xiao Jing,Zhenwei Zhu,Hongliang Li,Xin Pei,Yoshua Bengio,Tong Che,Hongyong Song

One of the greatest challenges of reinforcement learning is efficient exploration, especially when training signals are sparse or deceptive. The main difficulty of exploration lies in the size and complexity of the state space, which makes simple approaches such as exhaustive search infeasible. Our work is based on two important observations. On one hand, modern computing platforms are extremely scalable in terms of number of computing nodes and cores, which can complete asynchronous and well load-balanced computational tasks very fast. On the other hand, Divide-and-Conquer is a commonly used technique in computer science to solve similar problems (such as SAT) of doing efficient search in extremely large state space. In this paper, we apply the idea of divide-and-conquer in the context of intelligent exploration. The resulting exploration scheme can be combined with various specific intrinsic rewards designed for the given task. In our exploration scheme, the learning algorithm can automatically divide the state space into regions, and each agent is assigned to explore one of these regions. All the agents run asynchronously and they can be deployed onto modern distributed computing platforms. Our experiments show that the proposed method is highly efficient and is able to achieve state-of-the-art results in many RL tasks such as MiniGrid and Vizdoom.

****

Surreal-GAN:Semi-Supervised Representation Learning via GAN for uncovering heterogeneous disease-related imaging patterns

Zhijian Yang,Junhao Wen,Christos Davatzikos

A plethora of machine learning methods have been applied to imaging data, enabling the construction of clinically relevant imaging signatures of neurological and neuropsychiatric diseases. Oftentimes, such methods don't explicitly model the heterogeneity of disease effects, or approach it via nonlinear models that are not interpretable. Moreover, unsupervised methods may parse heterogeneity that is driven by nuisance confounding factors that affect brain structure or function, rather than heterogeneity relevant to a pathology of interest. On the other ha nd, semi-supervised clustering methods seek to derive a dichotomous subtype membership, ignoring the truth that disease heterogeneity spatially and temporally extends along a continuum. To address the aforementioned limitations, herein, we propose a novel method, termed Surreal-GAN (Semi-SUpeRvised ReprEsentAtion Learning via GAN). Using cross-sectional imaging data, Surreal-GAN dissects underlying disease-related heterogeneity under the principle of semi-supervised clustering (cluster mappings from normal control to patient), proposes a continuously dimensional representation, and infers the disease severity of patients at individual level along each dimension. The model first learns a transformation function from normal control (CN) domain to the patient (PT) domain with latent variables controlling transformation directions. An inverse mapping function together with regularization on function continuity, pattern orthogonality and monotonicity was also imposed to make sure that the transformation function captures necessarily meaningful imaging patterns with clinical significance. We first validated the model through extensive semi-synthetic experiments, and then demonstrate its potential in capturing biologically plausible imaging patterns in Alzheimer's disease (AD).

****

Adversarial Robustness as a Prior for Learned Representations

Logan Engstrom,Andrew Ilyas,Shibani Santurkar,Dimitris Tsipras,Brandon Tran,Aleksander Madry

An common goal in deep learning is to learn versatile, high-level feature representations of input data. However, standard networks' representations seem to possess shortcomings that, as we illustrate, prevent them from fully realizing this goal. In this work, we show that robust optimization can be re-cast as a tool for enforcing priors on the features learned by deep neural networks. It turns out that representations learned by robust models address the aforementioned shortcomings and make significant progress towards learning a high-level encoding of inputs. In particular, these representations are approximately invertible, while allowing for direct visualization and manipulation of salient input features. More broadly, our results indicate adversarial robustness as a promising avenue for improving learned representations.

****

SURF: Semi-supervised Reward Learning with Data Augmentation for Feedback-efficient Preference-based Reinforcement Learning

Jongjin Park,Younggyo Seo,Jinwoo Shin,Honglak Lee,Pieter Abbeel,Kimin Lee

Preference-based reinforcement learning (RL) has shown potential for teaching agents to perform the target tasks without a costly, pre-defined reward function by learning the reward with a supervisor’s preference between the two agent behaviors. However, preference-based learning often requires a large amount of human feedback, making it difficult to apply this approach to various applications. This data-efficiency problem, on the other hand, has been typically addressed by using unlabeled samples or data augmentation techniques in the context of supervised learning. Motivated by the recent success of these approaches, we present SURF, a semi-supervised reward learning framework that utilizes a large amount of unlabeled samples with data augmentation. In order to leverage unlabeled samples for reward learning, we infer pseudo-labels of the unlabeled samples based on the confidence of the preference predictor. To further improve the label-efficiency of reward learning, we introduce a new data augmentation that temporally crops consecutive subsequences from the original behaviors. Our experiments demonstrate that our approach significantly improves the feedback-efficiency of the state-of-the-art preference-based method on a variety of locomotion and robotic manipulation tasks.

****

Convergent Graph Solvers

Junyoung Park,Jinhyun Choo,Jinkyoo Park

We propose the convergent graph solver (CGS), a deep learning method that learns iterative mappings to predict the properties of a graph system at its stationary state (fixed point) with guaranteed convergence. The forward propagation of CGS proceeds in three steps: (1) constructing the input-dependent linear contracti

ng iterative maps, (2) computing the fixed points of the iterative maps, and (3) decoding the fixed points to estimate the properties. The contractivity of the constructed linear maps guarantees the existence and uniqueness of the fixed points following the Banach fixed point theorem. To train CGS efficiently, we also derive a tractable analytical expression for its gradient by leveraging the implicit function theorem. We evaluate the performance of CGS by applying it to various network-analytic and graph benchmark problems. The results indicate that CGS has competitive capabilities for predicting the stationary properties of graph systems, irrespective of whether the target systems are linear or non-linear. CGS also shows high performance for graph classification problems where the existence or the meaning of a fixed point is hard to be clearly defined, which highlights the potential of CGS as a general graph neural network architecture.
****

Noise-Contrastive Variational Information Bottleneck Networks
Jannik Schmitt,Stefan Roth
While deep neural networks for classification have shown impressive predictive performance, e.g. in image classification, they generally tend to be overconfident. We start from the observation that popular methods for reducing overconfidence by regularizing the distribution of outputs or intermediate variables achieve better calibration by sacrificing the separability of correct and incorrect predictions, another important facet of uncertainty estimation. To circumvent this, we propose a novel method that builds upon the distributional alignment of the variational information bottleneck and encourages assigning lower confidence to samples from the latent prior. Our experiments show that this simultaneously improves prediction accuracy and calibration compared to a multitude of output regularization methods without impacting the uncertainty-based separability in multiple classification settings, including under distributional shift.
****

Spread Spurious Attribute: Improving Worst-group Accuracy with Spurious Attribute Estimation
Junhyun Nam,Jaehyung Kim,Jaeho Lee,Jinwoo Shin
The paradigm of worst-group loss minimization has shown its promise in avoiding to learn spurious correlations, but requires costly additional supervision on spurious attributes. To resolve this, recent works focus on developing weaker forms of supervision---e.g., hyperparameters discovered with a small number of validation samples with spurious attribute annotation---but none of the methods retain comparable performance to methods using full supervision on the spurious attribute. In this paper, instead of searching for weaker supervisions, we ask: Given access to a fixed number of samples with spurious attribute annotations, what is the best achievable worst-group loss if we ''fully exploit'' them? To this end, we propose a pseudo-attribute-based algorithm, coined Spread Spurious Attribute (SSA), for improving the worst-group accuracy. In particular, we leverage samples both with and without spurious attribute annotations to train a model to predict the spurious attribute, then use the pseudo-attribute predicted by the trained model as supervision on the spurious attribute to train a new robust model having minimal worst-group loss. Our experiments on various benchmark datasets show that our algorithm consistently outperforms the baseline methods using the same number of validation samples with spurious attribute annotations. We also demonstrate that the proposed SSA can achieve comparable performances to methods using full (100%) spurious attribute supervision, by using a much smaller number of annotated samples---from 0.6% and up to 1.5%, depending on the dataset.
****

On Reward Maximization and Distribution Matching for Fine-Tuning Language Models
Tomasz Korbak,Hady Elsahar,Germán Kruszewski,Marc Dymetman
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, Dist

ribution 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 intimate 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.

****

Building the Building Blocks: From Simplification to Winning Trees in Genetic Programming

Lucija Planini■,Marko ■urasevi■,Stjepan Picek,Domagoj Jakobovic

Genetic Programming (GP) represents a powerful paradigm in diverse real-world applications. While GP can reach optimal (or at least ``good-enough'') solutions for many problems, such solutions are not without deficiencies. A frequent issue stems from the representation perspective where GP evolves solutions that contain unnecessary parts, known as program bloat.

This paper first investigates a combination of deterministic and random simplification to simplify the solutions while having a (relatively) small influence on the solution fitness. Afterward, we use the solutions to extract their subtrees, which we denote as winning trees. The winning trees can be used to initialize the population for the new GP run and result in improved convergence and fitness, provided some conditions on the size of solutions and winning trees are fulfilled. To experimentally validate our approach, we consider several synthetic benchmark problems and real-world symbolic regression problems.

****

The magnitude vector of images

Michael F Adamer,Leslie O'Bray,Edward De Brouwer,Bastian Rieck,Karsten Borgwardt

The magnitude of a finite metric space is a recently-introduced invariant quantity. Despite beneficial theoretical and practical properties, such as a general utility for outlier detection, and a close connection to Laplace radial basis kernels, magnitude has received little attention by the machine learning community so far. In this work, we investigate the properties of magnitude on individual images, with each image forming its own metric space. We show that the known properties of outlier detection translate to edge detection in images and we give supporting theoretical justifications. In addition, we provide a proof of concept of its utility by using a novel magnitude layer to defend against adversarial attacks. Since naive magnitude calculations may be computationally prohibitive, we introduce an algorithm that leverages the regular structure of images to dramatically reduce the computational cost.

****

Quasi-potential theory for escape problem: Quantitative sharpness effect on SGD's escape from local minima

Hikaru Ibayashi,Masaaki Imaizumi

We develop a quantitative theory on the escape problem of stochastic gradient descent (SGD) and investigate the effect of the sharpness of loss surfaces on escape. Deep learning has achieved tremendous success in various domains, however, it has opened up theoretical problems. For instance, it is still an ongoing question as to why an SGD can find solutions that generalize well over non-convex loss surfaces. An approach to explain this phenomenon is the escape problem, which investigates how efficiently the SGD escapes from local minima. In this paper, we develop a novel theoretical framework for the escape problem using ``quasi-potential," the notion defined in a fundamental theory of stochastic dynamical systems. We show that quasi-potential theory can handle the geometric property of loss surfaces and a covariance structure of gradient noise in a unified manner thr ough an eigenvalue argument, while previous research studied them separately. Our theoretical results imply that sharpness contributes to slowing down escape, but the SGD’s noise structure cancels the effect, which ends up exponentially accelerating its escape. We also conduct experiments to empirically validate our theory using neural networks trained with real data.

****

Learning Scenario Representation for Solving Two-stage Stochastic Integer Programs

Yaoxin Wu,Wen Song,Zhiguang Cao,Jie Zhang

Many practical combinatorial optimization problems under uncertainty can be modeled as stochastic integer programs (SIPs), which are extremely challenging to solve due to the high complexity. To solve two-stage SIPs efficiently, we propose a conditional variational autoencoder (CVAE) based method to learn scenario representation for a class of SIP instances. Specifically, we design a graph convolutional network based encoder to embed each scenario with the deterministic part of its instance (i.e. context) into a low-dimensional latent space, from which a decoder reconstructs the scenario from its latent representation conditioned on the context. Such a design effectively captures the dependencies of the scenarios on their corresponding instances. We apply the trained encoder to two tasks in typical SIP solving, i.e. scenario reduction and objective prediction. Experiments on two SIP problems show that the learned latent representation significantly boosts the solving performance to attain high-quality solutions in short computational time, and generalizes fairly well to problems of larger sizes or with more scenarios.

****

Generalization Through the Lens of Leave-One-Out Error

Gregor Bachmann,Thomas Hofmann,Aurelien Lucchi

Despite the tremendous empirical success of deep learning models to solve various learning tasks, our theoretical understanding of their generalization ability is very limited. Classical generalization bounds based on tools such as the VC dimension or Rademacher complexity, are so far unsuitable for deep models and it is doubtful that these techniques can yield tight bounds even in the most idealistic settings~\citep{nagarajan2019uniform}. In this work, we instead revisit the concept of leave-one-out (LOO) error to measure the generalization ability of deep models in the so-called kernel regime. While popular in statistics, the LOO error has been largely overlooked in the context of deep learning. By building upon the recently established connection between neural networks and kernel learning, we leverage the closed-form expression for the leave-one-out error, giving us access to an efficient proxy for the test error. We show both theoretically and empirically that the leave-one-out error is capable of capturing various phenomena in generalization theory, such as double descent, random labels or transfer learning.

Our work therefore demonstrates that the leave-one-out error provides a tractable way to estimate the generalization ability of deep neural networks in the kernel regime, opening the door to potential, new research directions in the field of generalization.

****

Containerized Distributed Value-Based Multi-Agent Reinforcement Learning

Siyang Wu,Tonghan Wang,Chenghao Li,Chongjie Zhang

Multi-agent reinforcement learning tasks put a high demand on the volume of training samples. Different from its single-agent counterpart, distributed value-based multi-agent reinforcement learning faces the unique challenges of demanding data transfer, inter-process communication management, and high requirement of exploration. We propose a containerized learning framework to solve these problems. We pack several environment instances, a local learner and buffer, and a carefully designed multi-queue manager which avoids blocking into a container. Local policies of each container are encouraged to be as diverse as possible, and only trajectories with highest priority are sent to a global learner. In this way, we achieve a scalable, time-efficient, and diverse distributed MARL learning framework with high system throughput. To own knowledge, our method is the first to solve the challenging Google Research Football full game $\mathtt{5\_v\_5}$. On the StarCraft II micromanagement benchmark, our method gets 4-18$\times$ better results compared to state-of-the-art non-distributed MARL algorithms.

****

Self-Supervised Inference in State-Space Models

David Ruhe,Patrick Forré

We perform approximate inference in state-space models with nonlinear state transitions. Without parameterizing a generative model, we apply Bayesian update formulas using a local linearity approximation parameterized by neural networks. It comes accompanied by a maximum likelihood objective that requires no supervision via uncorrupt observations or ground truth latent states. The optimization backpropagates through a recursion similar to the classical Kalman filter and smoother. Additionally, using an approximate conditional independence, we can perform smoothing without having to parameterize a separate model. In scientific applications, domain knowledge can give a linear approximation of the latent transition maps, which we can easily incorporate into our model. Usage of such domain knowledge is reflected in excellent results (despite our model's simplicity) on the chaotic Lorenz system compared to fully supervised and variational inference methods. Finally, we show competitive results on an audio denoising experiment.

****

Few-shot graph link prediction with domain adaptation

Hao Zhu,Mahashweta Das,Mangesh Bendre,Fei Wang,Hao Yang,Soha Hassoun

Real world link prediction problem often deals with data coming from multiple imbalanced domains. Similar problems in computer vision are often referred to as Few-Shot Learning (FSL) problems. However, for graph link prediction, this problem has rarely been addressed and explored. In this work, we propose an adversarial training based modification to the current state-of-the-arts link prediction method to solve this problem. We introduce a domain discriminator on pairs of graph-level embedding. We then use the discriminator to improve the model in an adversarial way, such that the graph embeddings generated by the model are domain agnostic. We test our proposal on 3 benchmark datasets. Our results demonstrate that, when domain differences exist, our method creates better graph embeddings that are more evenly distributed across domains and generates better prediction outcomes.

****

Improved Generalization Bound for Deep Neural Networks Using Geometric Functional Analysis

Phani raj Chinnalingu,Rajarshi Banerjee

Understanding how a neural network behaves in multiple domains is the key to further its explainability, generalizability, and robustness. In this paper, we prove a novel generalization bound using the fundamental concepts of geometric functional analysis. Specifically, by leveraging the covering number of the training dataset and applying certain geometric inequalities we show that a sharp bound can be obtained. To the best of our knowledge this is the first approach which utilizes covering numbers to estimate such generalization bounds.

****

$G^3$: Representation Learning and Generation for Geometric Graphs

Han Huang,Stefan C Schonsheck,Rongjie Lai,Jie Chen

A geometric graph is a graph equipped with geometric information (i.e., node coordinates). A notable example is molecular graphs, where the combinatorial bonding is supplement with atomic coordinates that determine the three-dimensional structure. This work proposes a generative model for geometric graphs, capitalizing on the complementary information of structure and geometry to learn the underlying distribution. The proposed model, Geometric Graph Generator (G$^3$), orchestrates graph neural networks and point cloud models in a nontrivial manner under an autoencoding framework. Additionally, we augment this framework with a normalizing flow so that one can effectively sample from the otherwise intractable latent space. G$^3$ can be used in computer-aided drug discovery, where seeking novel and optimal molecular structures is critical. As a representation learning ap proach, the interaction of the graph structure and the geometric point cloud also improve significantly the performance of downstream tasks, such as molecular property prediction. We conduct a comprehensive set of experiments to demonstrate that G^3 learns more accurately the distribution of given molecules and helps identify novel molecules with better properties of interest.

****

What can multi-cloud configuration learn from AutoML?

Malgorzata Lazuka,Thomas Parnell,Andreea Anghel,Haralampos Pozidis

Multi-cloud computing has become increasingly popular with enterprises looking to avoid vendor lock-in. While most cloud providers offer similar functionality, they may differ significantly in terms of performance and/or cost. A customer looking to benefit from such differences will naturally want to solve the multi-cloud configuration problem: given a workload, which cloud provider should be chosen and how should its nodes be configured in order to minimize runtime or cost? In this work, we consider this multi-cloud optimization problem and publish a new offline benchmark dataset, MOCCA, comprising 60 different multi-cloud configuration tasks across 3 public cloud providers, to enable further research in this area. Furthermore, we identify an analogy between multi-cloud configuration and the selection-configuration problems that are commonly studied in the automated machine learning (AutoML) field. Inspired by this connection, we propose an algorithm for solving multi-cloud configuration, CloudBandit (CB). It treats the outer problem of cloud provider selection as a best-arm identification problem, in which each arm pull corresponds to running an arbitrary black-box optimizer on the inner problem of node configuration. Extensive experiments on MOCCA indicate that CB achieves (a) significantly lower regret relative to its component black-box optimizers and (b) competitive or lower regret relative to state-of-the-art AutoML methods, whilst also being cheaper and faster.

****

On the Role of Neural Collapse in Transfer Learning

Tomer Galanti,András György,Marcus Hutter

We study the ability of foundation models to learn representations for classification that are transferable to new, unseen classes. Recent results in the literature show that representations learned by a single classifier over many classes are competitive on few-shot learning problems with representations learned by special-purpose algorithms designed for such problems. In this paper, we provide an explanation for this behavior based on the recently observed phenomenon that the features learned by overparameterized classification networks show an interesting clustering property, called neural collapse. We demonstrate both theoretically and empirically that neural collapse generalizes to new samples from the training classes, and -- more importantly -- to new classes as well, allowing foundation models to provide feature maps that work well in transfer learning and, specifically, in the few-shot setting.

****

Continuous Deep Q-Learning in Optimal Control Problems: Normalized Advantage Functions Analysis

Anton Plaksin,Stepan Martyanov

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 time-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.

****

EVaDE : Event-Based Variational Thompson Sampling for Model-Based Reinforcement Learning

Siddharth Aravindan,Dixant Mittal,Wee Sun Lee

Posterior Sampling for Reinforcement Learning (PSRL) is a well-known algorithm that augments model-based reinforcement learning (MBRL) algorithms with Thompson sampling. PSRL maintains posterior distributions of the environment transition dynamics and the reward function to procure posterior samples that are used to generate data for training the controller. Maintaining posterior distributions over all possible transition and reward functions for tasks with high dimensional state and action spaces is intractable. Recent works show that dropout used in conjunction with neural networks induce variational distributions that can approximate these posteriors. In this paper, we propose Event-based Variational Distributions for Exploration (EVaDE), variational distributions that are useful for MBRL, especially when the underlying domain is object-based. We leverage the general domain knowledge of object-based domains to design three types of event-based convolutional layers to direct exploration, namely the noisy event interaction layer, the noisy event weighting layer and the noisy event translation layer respectively. These layers rely on Gaussian dropouts and are inserted in between the layers of the deep neural network model to help facilitate variational Thompson sampling. We empirically show the effectiveness of EVaDE equipped Simulated Policy Learning (SimPLe) on a randomly selected suite of Atari games, where the number of agent environment interactions is limited to 100K.

****

Discovering Classification Rules for Interpretable Learning with Linear Programming

Hakan Akyuz,Ilker Birbil

Rules embody a set of if-then statements which include one or more conditions to classify a subset of samples in a dataset. In various applications such classification rules are considered to be interpretable by the decision makers. We introduce two new algorithms for interpretability and learning. Both algorithms take advantage of linear programming, and hence, they are scalable to large data sets. The first algorithm extracts rules for interpretation of trained models that are based on tree/rule ensembles. The second algorithm generates a set of classification rules through a column generation approach. The proposed algorithms return a set of rules along with their optimal weights indicating the importance of each rule for classification. Moreover, our algorithms allow assigning cost coefficients, which could relate to different attributes of the rules, such as; rule lengths, estimator weights, number of false negatives, and so on. Thus, the decision makers can adjust these coefficients to divert the training process and obtain a set of rules that are more appealing for their needs. We have tested the performances of both algorithms on a collection of datasets and presented a case study to elaborate on optimal rule weights. Our results show that a good compromise between interpretability and accuracy can be obtained by the proposed algorithms.

****

ParaDiS: Parallelly Distributable Slimmable Neural Networks

Alexey Ozerov,Anne Lambert,Suresh Kirthi Kumaraswamy

When several limited power devices are available, one of the most efficient ways to make profit of these resources, while reducing the processing latency and communication load, is to run in parallel several neural sub-networks and to fuse the result at the end of processing. However, such a combination of sub-networks must be trained specifically for each particular configuration of devices (characterized by number of devices and their capacities) which may vary over different model deployments and even within the same deployment. In this work we introduce parallelly distributable slimmable (ParaDiS) neural networks that are splittable in parallel among various device configurations without retraining. While inspired by slimmable networks allowing instant adaptation to resources on just one device, ParaDiS networks consist of several multi-device distributable configurations or switches that strongly share the parameters between them. We evaluate ParaDiS framework on MobileNet v1 and ResNet-50 architectures on ImageNet classification task. We show that ParaDiS switches achieve similar or better accuracy than the individual models, i.e., distributed models of the same structure tr ained individually. Moreover, we show that, as compared to universally slimmable networks that are not distributable, the accuracy of distributable ParaDiS switches either does not drop at all or drops by a maximum of 1 % only in the worst cases.

****

Complex Locomotion Skill Learning via Differentiable Physics

Jiancheng Liu,Yu Fang,Mingrui Zhang,Jiasheng Zhang,Yidong Ma,Minchen Li,Yuanming Hu,Chenfanfu Jiang,Tiantian Liu

Differentiable physics enables efficient gradient-based optimizations of neural network (NN) controllers. However, existing work typically only delivers NN controllers with limited capability and generalizability. We present a practical learning framework that outputs unified NN controllers capable of tasks with significantly improved complexity and diversity. To systematically improve training robustness and efficiency, we investigated a suite of improvements over the baseline approach, including periodic activation functions, and tailored loss functions. In addition, we find our adoption of batching and a modified Adam optimizer effective in training complex locomotion tasks. We evaluate our framework on differentiable mass-spring and material point method (MPM) simulations, with challenging locomotion tasks and multiple robot designs. Experiments show that our learning framework, based on differentiable physics, delivers better results than reinforcement learning and converges much faster. We demonstrate that users can interactively control soft robot locomotion and switch among multiple goals with specified velocity, height, and direction instructions using a unified NN controller trained in our system.

****

Information-theoretic Online Memory Selection for Continual Learning

Shengyang Sun,Daniele Calandriello,Huiyi Hu,Ang Li,Michalis Titsias

A challenging problem in task-free continual learning is the online selection of a representative replay memory from data streams. In this work, we investigate the online memory selection problem from an information-theoretic perspective. To gather the most information, we propose the \textit{surprise} and the \textit{learnability} criteria to pick informative points and to avoid outliers. We present a Bayesian model to compute the criteria efficiently by exploiting rank-one matrix structures. We demonstrate that these criteria encourage selecting informative points in a greedy algorithm for online memory selection. Furthermore, by identifying the importance of \textit{the timing to update the memory}, we introduce a stochastic information-theoretic reservoir sampler (InfoRS), which conducts sampling among selective points with high information. Compared to reservoir sampling, InfoRS demonstrates improved robustness against data imbalance. Finally, empirical performances over continual learning benchmarks manifest its efficiency and efficacy.

****

Dealing with Non-Stationarity in MARL via Trust-Region Decomposition

Wenhao Li,Xiangfeng Wang,Bo Jin,Junjie Sheng,Hongyuan Zha

Non-stationarity is one thorny issue in cooperative multi-agent reinforcement learning (MARL). One of the reasons is the policy changes of agents during the learning process. Some existing works have discussed various consequences caused by non-stationarity with several kinds of measurement indicators. This makes the objectives or goals of existing algorithms are inevitably inconsistent and disparate. In this paper, we introduce a novel notion, the δ-$stationarity$ measurement, to explicitly measure the non-stationarity of a policy sequence, which can be further proved to be bounded by the KL-divergence of consecutive joint policies. A straightforward but highly non-trivial way is to control the joint policies' divergence, which is difficult to estimate accurately by imposing the trust-region constraint on the joint policy. Although it has lower computational complexity to decompose the joint policy and impose trust-region constraints on the factorized policies, simple policy factorization like mean-field approximation will lead to more considerable policy divergence, which can be considered as the trust-region decomposition dilemma. We model the joint policy as a pairwise Markov random field and propose a trust-region decomposition network (TRD-Net) ba sed on message passing to estimate the joint policy divergence more accurately. The Multi-Agent Mirror descent policy algorithm with Trust region decomposition, called MAMT, is established by adjusting the trust-region of the local policies adaptively in an end-to-end manner. MAMT can approximately constrain the consecutive joint policies' divergence to satisfy δ-stationarity and alleviate the non-stationarity problem. Our method can bring noticeable and stable performance improvement compared with baselines in cooperative tasks of different complexity.

****

Information Bottleneck: Exact Analysis of (Quantized) Neural Networks

Stephan Sloth Lorenzen,Christian Igel,Mads Nielsen

The information bottleneck (IB) principle has been suggested as a way to analyze deep neural networks. The learning dynamics are studied by inspecting the mutual information (MI) between the hidden layers and the input and output. Notably, separate fitting and compression phases during training have been reported. This led to some controversy including claims that the observations are not reproducible and strongly dependent on the type of activation function used as well as on the way the MI is estimated. Our study confirms that different ways of binning when computing the MI lead to qualitatively different results, either supporting or refusing IB conjectures.

To resolve the controversy, we study the IB principle in settings where MI is non-trivial and can be computed exactly. We monitor the dynamics of quantized neural networks, that is, we discretize the whole deep learning system so that no approximation is required when computing the MI. This allows us to quantify the information flow without measurement errors.

In this setting, we observed a fitting phase for all layers and a compression phase for the output layer in all experiments; the compression in the hidden layers was dependent on the type of activation function. Our study shows that the initial IB results were not artifacts of binning when computing the MI. However, the critical claim that the compression phase may not be observed for some networks also holds true.

****

HyperCGAN: Text-to-Image Synthesis with HyperNet-Modulated Conditional Generative Adversarial Networks

Kilichbek Haydarov,Aashiq Muhamed,Jovana Lazarevic,Ivan Skorokhodov,Mohamed Elhoseiny

We present HyperCGAN: a conceptually simple and general approach for text-to-image synthesis that uses hypernetworks to condition a GAN model on text. In our setting, the generator and the discriminator weights are controlled by their corresponding hypernetworks, which modulate weight parameters based on the provided text query. We explore different mechanisms to modulate the layers depending on the underlying architecture of a target network and the structure of the conditioning variable. Our method enjoys high flexibility, and we test it in two scenarios: traditional image generation (on top of StyleGAN2) and continuous image generation (on top of INR-GAN). To the best of our knowledge, our work is the first one which explores text-controllable continuous image generation. In both cases, hypernetwork-based conditioning achieves state-of-the-art performance in terms of modern text-to-image evaluation measures and human studies on CUB 256^2, COCO 256^2, and ArtEmis 256^2 datasets.

****

Learnability of convolutional neural networks for infinite dimensional input via mixed and anisotropic smoothness

Sho Okumoto,Taiji Suzuki

Among a wide range of success of deep learning, convolutional neural networks have been extensively utilized in several tasks such as speech recognition, image processing, and natural language processing, which require inputs with large dimensions.

Several studies have investigated function estimation capability of deep learning, but most of them have assumed that the dimensionality of the input is much smaller than the sample size.

However, for typical data in applications such as those handled by the convolutional neural networks described above,
the dimensionality of inputs is relatively high or even infinite.
In this paper, we investigate the approximation and estimation errors of the (dilated) convolutional neural networks when the input is infinite dimensional.
Although the approximation and estimation errors of neural networks are affected by the curse of dimensionality in the existing analyses for typical function spaces such as the \Holder and Besov spaces, we show that, by considering anisotropic smoothness, they can alleviate exponential dependency on the dimensionality but they only depend on the smoothness of the target functions.
Our theoretical analysis supports the great practical success of convolutional networks.
Furthermore, we show that the dilated convolution is advantageous when the smoothness of the target function has a sparse structure.
****
Objective Evaluation of Deep Visual Interpretations on Time Series Data
Christoffer Löffler,Wei-Cheng Lai,Lukas M Schmidt,Dario Zanca,Bjoern Eskofier,Christopher Mutschler
The correct interpretation and understanding of deep learning models is essential in many applications.
(Explanatory) visual interpretation approaches for image and natural language processing allow domain experts to validate and understand almost any deep learning model. However, they fall short when generalizing to arbitrary time series data that is less intuitive and more diverse. Whether a visualization explains the true reasoning or captures the real features is more difficult to judge. Hence, instead of blind trust we need an objective evaluation to obtain reliable quality metrics. This paper proposes a framework of six orthogonal quality metrics for gradient- or perturbation-based post-hoc visual interpretation methods designed for time series classification and segmentation tasks. This comprehensive set is either based on "human perception" or on "functional properties". An extensive experimental study includes commonly used neural network architectures for time series and nine visual interpretation methods. We evaluate the visual interpretation methods with diverse datasets from the UCR repository as well another complex real-world dataset. We show that none of the existing methods consistently outperforms any of the others on all metrics while some of them are ahead in either functional or human-based metrics. Our results allow experts to make an informed choice of suitable visualization techniques for the model and task at hand.
****
GLASS: GNN with Labeling Tricks for Subgraph Representation Learning
Xiyuan Wang,Muhan Zhang
Despite the remarkable achievements of Graph Neural Networks (GNNs) on graph representation learning, few works have tried to use them to predict properties of subgraphs in the whole graph. The existing state-of-the-art method SubGNN introduces an overly complicated subgraph-level GNN model which synthesizes three artificial channels each of which has two carefully designed subgraph-level message passing modules, yet only slightly outperforms a plain GNN which performs node-level message passing and then pools node embeddings within the subgraph. By analyzing SubGNN and plain GNNs, we find that the key for subgraph representation learning might be to distinguish nodes inside and outside the subgraph. With this insight, we propose an expressive and scalable labeling trick, namely max-zero-one, to enhance plain GNNs for subgraph tasks. The resulting model is called GLASS (GNN with LAbeling trickS for Subgraph). We theoretically characterize GLASS's expressive power. Compared with SubGNN, GLASS is more expressive, more scalable, and easier to implement. Experiments on eight benchmark datasets show that GLASS outperforms the strongest baseline by 14.8% on average. And ablation analysis shows that our max-zero-one labeling trick can boost the performance of a plain GNN by up to 105% in maximum, which illustrates the effectiveness of labeling trick on subgraph tasks. Furthermore, training a GLASS model only takes 37% time needed for a SubGNN on average.
****

MoReL: Multi-omics Relational Learning

Arman Hasanzadeh,Ehsan Hajiramezanali,Nick Duffield,Xiaoning Qian

Multi-omics data analysis has the potential to discover hidden molecular interactions, revealing potential regulatory and/or signal transduction pathways for cellular processes of interest when studying life and disease systems. One of critical challenges when dealing with real-world multi-omics data is that they may manifest heterogeneous structures and data quality as often existing data may be collected from different subjects under different conditions for each type of omics data. We propose a novel deep Bayesian generative model to efficiently infer a multi-partite graph encoding molecular interactions across such heterogeneous views, using a fused Gromov-Wasserstein (FGW) regularization between latent representations of corresponding views for integrative analysis. With such an optimal transport regularization in the deep Bayesian generative model, it not only allows incorporating view-specific side information, either with graph-structured or unstructured data in different views, but also increases the model flexibility with the distribution-based regularization. This allows efficient alignment of heterogeneous latent variable distributions to derive reliable interaction predictions compared to the existing point-based graph embedding methods. Our experiments on several real-world datasets demonstrate enhanced performance of MoReL in inferring meaningful interactions compared to existing baselines.

****

SPLID: Self-Imitation Policy Learning through Iterative Distillation

Zhihan Liu,Hao Sun,Bolei Zhou

Goal-Conditioned continuous control tasks remain challenging due to the sparse reward signals. To address this issue, many relabelling methods like Hindsight Experience Replay have been developed and bring significant improvement. Though relabelling methods provide an alternative to an expert demonstration, the majority of the relabelled data are not optimal. If we can improve the quality of the relabelled data, the sample efficiency, as well as the agent performance, should be improved. To this end, we propose a novel meta-algorithm Self-Imitation Policy Learning through Iterative Distillation (SPLID) which relies on the concept of δ-distilled policy to iteratively level up the quality of the target data and agent mimics from the relabeled target data. Under certain assumptions, we show that SPLID has good theoretical properties of performance improvement and local convergence guarantee. Specifically, in the deterministic environment, we develop a practical implementation of SPLID, which imposes δ-distilled policy by discriminating First Hit Time (FHT). Experiments show that SPLID outperforms previous Goal-Conditioned RL methods with a substantial margin.

****

Provable Learning-based Algorithm For Sparse Recovery

Xinshi Chen,Haoran Sun,Le Song

Recovering sparse parameters from observational data is a fundamental problem in machine learning with wide applications. Many classic algorithms can solve this problem with theoretical guarantees, but their performances rely on choosing the correct hyperparameters. Besides, hand-designed algorithms do not fully exploit the particular problem distribution of interest. In this work, we propose a deep learning method for algorithm learning called PLISA (Provable Learning-based Iterative Sparse recovery Algorithm). PLISA is designed by unrolling a classic path-following algorithm for sparse recovery, with some components being more flexible and learnable. We theoretically show the improved recovery accuracy achievable by PLISA. Furthermore, we analyze the empirical Rademacher complexity of PLISA to characterize its generalization ability to solve new problems outside the training set. This paper contains novel theoretical contributions to the area of learning-based algorithms in the sense that (i) PLISA is generically applicable to a broad class of sparse estimation problems, (ii) generalization analysis has received less attention so far, and (iii) our analysis makes novel connections between the generalization ability and algorithmic properties such as stability and convergence of the unrolled algorithm, which leads to a tighter bound that can explain the empirical observations. The techniques could potentially be applied to analyze other learning-based algorithms in the literature.

****

Multi-Objective Online Learning

Jiyan Jiang,Wenpeng Zhang,Shiji Zhou,Lihong Gu,Xiaodong Zeng,Wenwu Zhu

This paper presents a systematic study of multi-objective online learning. We first formulate the framework of Multi-Objective Online Convex Optimization, which encompasses a novel multi-objective dynamic regret in the unconstrained max-min form. We show that it is equivalent to the regret commonly used in the zero-order multi-objective bandit setting and overcomes the problem that the latter is hard to optimize via first-order gradient-based methods. Then we propose the Online Mirror Multiple Descent algorithm with two variants, which computes the composite gradient using either the vanilla min-norm solver or a newly designed L_1-regularized min-norm solver. We further derive regret bounds of both variants and show that the L_1-regularized variant enjoys a lower bound. Extensive experiments demonstrate the effectiveness of the proposed algorithm and verify the theoretical advantage of the L_1-regularized variant.

****

GIR Framework: Learning Graph Positional Embeddings with Anchor Indication and Path Encoding

Yuheng Lu,Jinpeng Chen,Chuxiong Sun,Jie Hu

The majority of existing graph neural networks (GNNs) following the message passing neural network (MPNN) pattern have limited power in capturing position information for a given node. To solve such problems, recent works exploit positioning nodes with selected anchors, mostly in a process that first explicitly assign distances information and then perform message passing encoding. However, this two-stage strategy may ignore potentially useful interaction between intermediate results of the distance computing and encoding stages. In this work, we propose a novel framework which follows the anchor-based idea and aims at conveying distance information implicitly along the MPNN message passing steps for encoding position information, node attributes, and graph structure in a more flexible way. Specifically, we first leverage a simple anchor indication strategy to enable the position-aware ability for well-designed MPNNs. Then, following this strategy, we propose the Graph Inference Representation (GIR) model, which acts as a generalization of MPNNs with a more specific propagation path design for position-aware scenarios. Meanwhile, we theoretically and empirically explore the ability of the proposed framework to get position-aware embeddings, and experimental results show that our proposed method generally outperforms previous position-aware GNN methods.

****

Defending Against Image Corruptions Through Adversarial Augmentations

Dan Andrei Calian,Florian Stimberg,Olivia Wiles,Sylvestre-Alvise Rebuffi,András György,Timothy A Mann,Sven Gowal

Modern neural networks excel at image classification, yet they remain vulnerable to common image corruptions such as blur, speckle noise or fog. Recent methods that focus on this problem, such as AugMix and DeepAugment, introduce defenses that operate in expectation over a distribution of image corruptions. In contrast, the literature on Lp-norm bounded perturbations focuses on defenses against worst-case corruptions. In this work, we reconcile both approaches by proposing AdversarialAugment, a technique which optimizes the parameters of image-to-image models to generate adversarially corrupted augmented images. We theoretically motivate our method and give sufficient conditions for the consistency of its idealized version as well as that of DeepAugment. Our classifiers improve upon the state-of-the-art on common image corruption benchmarks conducted in expectation on CIFAR-10-C and improve worst-case performance against Lp-norm bounded perturbations on both CIFAR-10 and ImageNet.

****

Evolving Neural Update Rules for Sequence Learning

Karol Gregor,Peter Conway Humphreys

We consider the problem of searching, end to end, for effective weight and activation update rules governing online learning of a recurrent network on problems of character sequence memorisation and prediction. We experiment with a number o

f functional forms and find that the performance depends on them significantly. We find update rules that allow us to scale to a much larger number of recurrent units and much longer sequence lengths than has been achieved with this approach previously. We also find that natural evolution strategies significantly outperforms meta-gradients on this problem, aligning with previous studies suggesting that such evolutionary strategies are more robust than gradient back-propagation over sequences with thousands(s) of steps.
****

Attacking deep networks with surrogate-based adversarial black-box methods is easy
Nicholas A. Lord,Romain Mueller,Luca Bertinetto
A recent line of work on black-box adversarial attacks has revived the use of transfer from surrogate models by integrating it into query-based search. However, we find that existing approaches of this type underperform their potential, and can be overly complicated besides. Here, we provide a short and simple algorithm which achieves state-of-the-art results through a search which uses the surrogate network's class-score gradients, with no need for other priors or heuristics. The guiding assumption of the algorithm is that the studied networks are in a fundamental sense learning similar functions, and that a transfer attack from one to the other should thus be fairly "easy". This assumption is validated by the extremely low query counts and failure rates achieved: e.g. an untargeted attack on a VGG-16 ImageNet network using a ResNet-152 as the surrogate yields a median query count of 6 at a success rate of 99.9%. Code is available at https://github.com/fiveai/GFCS.
****

Distributed Zeroth-Order Optimization: Convergence Rates That Match Centralized Counterpart
Deming Yuan,Lei Wang,Alexandre Proutiere,Guodong Shi
Zeroth-order optimization has become increasingly important in complex optimization and machine learning when cost functions are impossible to be described in closed analytical forms. The key idea of zeroth-order optimization lies in the ability for a learner to build gradient estimates by queries sent to the cost function, and then traditional gradient descent algorithms can be executed with gradients replaced by the estimates. For optimization of large-scale multi-agent systems with decentralized data and costs, zeroth-order optimization can continue to be utilized to develop scalable and distributed zeroth-order algorithms. It is important to understand the trend in performance transitioning from centralized to distributed zeroth-order algorithms in terms of convergence rates, especially for multi-agent systems with time-varying communication networks. In this paper, we establish a series of convergence rates for distributed zeroth-order subgradient algorithms under both one-point and two-point zeroth-order oracles. Apart from the additional node-to-node communication cost in distributed algorithms, the established rates in convergence are shown to match their centralized counterpart. We also propose a multi-stage distributed zeroth-order algorithm that better utilizes the learning rates, reduces the computational complexity,   and attains even faster convergence rates for compact decision set.
****

Hypothesis Driven Coordinate Ascent for Reinforcement Learning
John Kenton Moore,Junier Oliva
This work develops a novel black box optimization technique for learning robust policies for stochastic environments. Through combining coordinate ascent with hypothesis testing, Hypothesis Driven Coordinate Ascent (HDCA) optimizes without computing or estimating gradients. The simplicity of this approach allows it to excel in a distributed setting; its implementation provides an interesting alternative to many state-of-the-art methods for common reinforcement learning environments. HDCA was evaluated on various problems from the MuJoCo physics simulator and OpenAI Gym framework, achieving equivalent or superior results to standard RL benchmarks.
****

Logic Pre-Training of Language Models

Siru Ouyang,Zhuosheng Zhang,hai zhao

Pre-trained language models (PrLMs) have been shown useful for enhancing a broad range of natural language understanding (NLU) tasks. However, the capacity for capturing logic relations in challenging NLU still remains a bottleneck even for state-of-the-art PrLM enhancement, which greatly stalled their reasoning abilities. Thus we propose logic pre-training of language models, leading to the logic reasoning ability equipped PrLM, \textsc{Prophet}. To let logic pre-training perform on a clear, accurate, and generalized knowledge basis, we introduce \textit{fact} instead of the plain language unit in previous PrLMs. The \textit{fact} is extracted through syntactic parsing in avoidance of unnecessary complex knowledge injection. Meanwhile, it enables training logic-aware models to be conducted on a more general language text. To explicitly guide the PrLM to capture logic relations, three pre-training objectives are introduced: 1) logical connectives masking to capture sentence-level logics, 2) logical structure completion to accurately capture facts from the original context, 3) logical path prediction on a logical graph to uncover global logic relationships among facts. We evaluate our model on a broad range of NLP and NLU tasks, including natural language inference, relation extraction, and machine reading comprehension with logical reasoning. Results show that the extracted fact and the newly introduced pre-training tasks can help \textsc{Prophet} achieve significant performance in all the downstream tasks, especially in logic reasoning related tasks.

****

Scaling Fair Learning to Hundreds of Intersectional Groups

Eric Zhao,De-An Huang,Hao Liu,Zhiding Yu,Anqi Liu,Olga Russakovsky,Anima Anandkumar

Bias mitigation algorithms aim to reduce the performance disparity between different protected groups. Existing techniques focus on settings where there is a small number of protected groups arising from a single protected attribute, such as skin color, gender or age. In real-world applications, however, there are multiple protected attributes yielding a large number of intersectional protected groups. These intersectional groups are particularly prone to severe underrepresentation in datasets. We conduct the first thorough empirical analysis of how existing bias mitigation methods scale to this setting, using large-scale datasets including the ImageNet People Subtree and CelebA. We find that as more protected attributes are introduced to a task, it becomes more important to leverage the protected attribute labels during training to promote fairness. We also find that the use of knowledge distillation, in conjunction with group-specific models, can help scale existing fair learning methods to hundreds of protected intersectional groups and reduce bias. We show on ImageNet's People Subtree that combining these insights can further reduce the bias amplification of fair learning algorithms by 15% ---a surprising reduction given that the dataset has 196 protected groups but fewer than 10% of the training dataset has protected attribute labels.

****

Zero-Cost Operation Scoring in Differentiable Architecture Search

Lichuan Xiang,■ukasz Dudziak,Mohamed S Abdelfattah,Thomas Chun Pong Chau,Nicholas Donald Lane,Hongkai Wen

Differentiable neural architecture search (NAS) has attracted significant attention in recent years due to its ability to quickly discover promising architectures of deep neural networks even in very large search spaces. Despite its success, many differentiable NAS methods lack robustness and may degenerate to trivial architectures with excessive parameter-free operations such as skip connections thus leading to inferior performance. In fact, selecting operations based on the magnitude of architectural parameters was recently proven to be fundamentally wrong, showcasing the need to rethink how operation scoring and selection occurs in differentiable NAS. To this end, we formalize and analyze a fundamental component of differentiable NAS: local "operation scoring" that occurs at each choice of operation.

When comparing existing operation scoring functions, we find that existing methods can be viewed as inexact proxies for accuracy.

We also find that existing methods perform poorly when analyzed empirically on NAS benchmarks. From this perspective, we introduce new training-free proxies to the context of differentiable NAS, and show that we can significantly speed up the search process while improving accuracy on multiple search spaces. We take inspiration from zero-cost proxies that were recently studied in the context of sample-based NAS but shown to degrade significantly for larger search spaces like DARTS. Our novel "perturbation-based zero-cost operation scoring" (Zero-Cost-PT) improves searching time and accuracy compared to the best available differentiable architecture search for many search space sizes, including very large ones. Specifically, we are able improve accuracy compared to the best current method (DARTS-PT) on the DARTS CNN search space while being over 40x faster (total searching time 25 minutes on a single GPU). Our code is available at: https://github.com/avail-upon-acceptance.

****

Experience Replay More When It's a Key Transition in Deep Reinforcement Learning

Youtian Guo,Qi Gao

We propose a experience replay mechanism in Deep Reinforcement Learning based on Add Noise to Noise (AN2N), which requires agent to replay more experience containing key state, abbreviated as Experience Replay More (ERM). In the AN2N algorithm, we refer to the states where exploring more as the key states. We found that how the transitions containing the key state participates in updating the policy and Q networks has a significant impact on the performance improvement of the deep reinforcement learning agent, and the problem of catastrophic forgetting in neural networks is further magnified in the AN2N algorithm. Therefore, we change the previous strategy of uniform sampling of experience transitions. We sample the transition used for experience replay according to whether the transition contains key states and whether it is the most recently generated, which is the core idea of the ERM algorithm. The experimental results show that this algorithm can significantly improve the performance of the agent. We combine the ERM algorithm with Deep Deterministic Policy Gradient (DDPG), Twin Delayed Deep Deterministic policy gradient (TD3) and Soft Actor-Critic (SAC), and evaluate algorithm on the suite of OpenAI gym tasks, SAC with ERM achieves a new state of the art, and DDPG with ERM can even exceed the average performance of SAC under certain random seeds, which is incredible.

****

Autoregressive Diffusion Models

Emiel Hoogeboom,Alexey A. Gritsenko,Jasmijn Bastings,Ben Poole,Rianne van den Berg,Tim Salimans

We introduce Autoregressive Diffusion Models (ARDMs), a model class encompassing and generalizing order-agnostic autoregressive models (Uria et al., 2014) and absorbing discrete diffusion (Austin et al., 2021), which we show are special cases of ARDMs under mild assumptions. ARDMs are simple to implement and easy to train. Unlike standard ARMs, they do not require causal masking of model representations, and can be trained using an efficient objective similar to modern probabilistic diffusion models that scales favourably to highly-dimensional data. At test time, ARDMs support parallel generation which can be adapted to fit any given generation budget. We find that ARDMs require significantly fewer steps than discrete diffusion models to attain the same performance. Finally, we apply ARDMs to lossless compression, and show that they are uniquely suited to this task. Contrary to existing approaches based on bits-back coding, ARDMs obtain compelling results not only on complete datasets, but also on compressing single data points. Moreover, this can be done using a modest number of network calls for (de)compression due to the model's adaptable parallel generation.

****

Contrastive Attraction and Contrastive Repulsion for Representation Learning

Huangjie Zheng,Xu Chen,Jiangchao Yao,Hongxia Yang,Chunyuan Li,Ya Zhang,Hao Zhang,Ivor Tsang,Jingren Zhou,Mingyuan Zhou

Contrastive learning (CL) methods effectively learn data representations without label supervision, where the encoder needs to contrast each positive sample over multiple negative samples via a one-vs-many softmax cross-entropy loss. By lev eraging large amounts of unlabeled image data, recent CL methods have achieved promising results when pretrained on ImageNet, a well-curated dataset with balanced image classes. However, they tend to yield worse performance when pretrained on images in the wild. In this paper, to further improve the performance of CL and enhance its robustness on uncurated datasets, we propose a doubly CL strategy that contrasts positive samples and negative ones within themselves separately. We realize this strategy with contrastive attraction and contrastive repulsion (CACR), which makes the query not only exert a greater force to attract more distant positive samples but also do so to repel closer negative samples. Theoretical analysis reveals that CACR generalizes CL's behavior by positive attraction and negative repulsion. It further considers the intra-contrastive relation within the positive and negative pairs to narrow the gap between the sampled and true distribution, which is important when datasets are less curated. Extensive large-scale experiments on standard vision tasks show that CACR not only consistently outperforms existing CL methods on benchmark datasets in representation learning, but also shows better robustness when generalized to pretrain on wild large image datasets.

****

Auto-scaling Vision Transformers without Training

Wuyang Chen,Wei Huang,Xianzhi Du,Xiaodan Song,Zhangyang Wang,Denny Zhou

This work targets automated designing and scaling of Vision Transformers (ViTs). The motivation comes from two pain spots: 1) the lack of efficient and principled methods for designing and scaling ViTs; 2) the tremendous computational cost of training ViT that is much heavier than its convolution counterpart. To tackle these issues, we propose As-ViT, an auto-scaling framework for ViTs without training, which automatically discovers and scales up ViTs in an efficient and principled manner. Specifically, we first design a "seed" ViT topology by leveraging a training-free search process. This extremely fast search is fulfilled by a comprehensive study of ViT's network complexity, yielding a strong Kendall-tau correlation with ground-truth accuracies. Second, starting from the "seed" topology, we automate the scaling rule for ViTs by growing widths/depths to different ViT layers. This results in a series of architectures with different numbers of parameters in a single run. Finally, based on the observation that ViTs can tolerate coarse tokenization in early training stages, we propose a progressive tokenization strategy to train ViTs faster and cheaper. As a unified framework, As-ViT achieves strong performance on classification (83.5% top1 on ImageNet-1k) and detection (52.7% mAP on COCO) without any manual crafting nor scaling of ViT architectures: the end-to-end model design and scaling process costs only 12 hours on one V100 GPU. Our code is available at https://github.com/VITA-Group/AsViT.

****

Network robustness as a mathematical property: training, evaluation and attack

Marco Casadio,Matthew L Daggitt,Ekaterina Komendantskaya,Wen Kokke,Robert Stewart

Neural networks are widely used in AI for their ability to detect general patterns in noisy data. Paradoxically, by default they are also known to not be particularly robust, i.e. moving a small distance in the input space can result in the network's output changing significantly.

Many methods for improving neural network robustness have been proposed recently. This growing body of research gave rise to numerous explicit or implicit notions of robustness. Connections between these notions are often subtle, and a systematic comparison of these different definitions was lacking in the literature.

In this paper we attempt to address this gap by performing an in-depth comparison of the different definitions of robustness, by analysing their relationships, assumptions, interpretability and verifiability.

By abstracting robustness as a stand-alone mathematical property, we are able to show that, having a choice of several definitions of robustness, one can combine them in a modular way when defining training modes, evaluation metrics, and attacks on neural networks.

We also perform experiments to compare the applicability and efficacy of different training methods for ensuring the network obeys these different definitions.

****

Fine-grained Differentiable Physics: A Yarn-level Model for Fabrics

Deshan Gong,Zhanxing Zhu,Andrew J. Bulpitt,He Wang

Differentiable physics modeling combines physics models with gradient-based learning to provide model explicability and data efficiency. It has been used to learn dynamics, solve inverse problems and facilitate design, and is at its inception of impact. Current successes have concentrated on general physics models such as rigid bodies, deformable sheets, etc, assuming relatively simple structures and forces. Their granularity is intrinsically coarse and therefore incapable of modelling complex physical phenomena. Fine-grained models are still to be developed to incorporate sophisticated material structures and force interactions with gradient-based learning. Following this motivation, we propose a new differentiable fabrics model for composite materials such as cloths, where we dive into the granularity of yarns and model individual yarn physics and yarn-to-yarn interactions. To this end, we propose several differentiable forces, whose counterparts in empirical physics are indifferentiable, to facilitate gradient-based learning. These forces, albeit applied to cloths, are ubiquitous in various physical systems. Through comprehensive evaluation and comparison, we demonstrate our model's $\textit{explicability}$ in learning meaningful physical parameters, $\textit{versatility}$ in incorporating complex physical structures and heterogeneous materials, $\textit{data-efficiency}$ in learning, and $\textit{high-fidelity}$ in capturing subtle dynamics.

****

Revisiting flow generative models for Out-of-distribution detection

Dihong Jiang,Sun Sun,Yaoliang Yu

Deep generative models have been widely used in practical applications such as the detection of out-of-distribution (OOD) data. In this work, we aim to re-examine the potential of generative flow models in OOD detection. We first propose a simple combination of univariate one-sample statistical test (e.g., Kolmogorov-Smirnov) and random projections in the latent space of flow models to perform OOD detection. Then, we propose a two-sample version of our test to account for imperfect flow models. Quite distinctly, our method does not pose parametric assumptions on OOD data and is capable of exploiting any flow model. Experimentally, firstly we confirm the efficacy of our method against state-of-the-art baselines through extensive experiments on several image datasets; secondly we investigate the relationship between model accuracy (e.g., the generation quality) and the OOD detection performance, and found surprisingly that they are not always positively correlated; and thirdly we show that detection in the latent space of flow models generally outperforms detection in the sample space across various OOD datasets, hence highlighting the benefits of training a flow model.

****

Distribution-Driven Disjoint Prediction Intervals for Deep Learning

Jaehak Cho,Jae Myung Kim,Sungyeob Han,Jungwoo Lee

This paper redefines prediction intervals (PIs) as the form of a union of disjoint intervals. PIs represent predictive uncertainty in the regression problem. Since previous PI methods assumed a single continuous PI (one lower and upper bound), it suffers from performance degradation in the uncertainty estimation when the conditional density function has multiple modes. This paper demonstrates that multimodality should be considered in regression uncertainty estimation. To address the issue, we propose a novel method that generates a union of disjoint PIs. Throughout UCI benchmark experiments, our method improves over current state-of-the-art uncertainty quantification methods, reducing an average PI width by over 27$\%$. Through qualitative experiments, we visualized that the multi-mode often exists in real-world datasets and why our method produces high-quality PIs compared to the previous PI.

****

MS^2-Transformer: An End-to-End Model for MS/MS-assisted Molecule Identification

Mengji Zhang,Yingce Xia,Nian Wu,Kun Qian,Jianyang Zeng

Mass spectrometry (MS) acts as an important technique for measuring the mass-to-

charge ratios of ions and identifying the chemical structures of unknown metabolites. Practically, tandem mass spectrometry (MS/MS), which couples multiple standard MS in series and outputs fine-grained spectrum with fragmental information, has been popularly used. Manually interpreting the MS/MS spectrum into the molecules (i.e., the simplified molecular-input line-entry system, SMILES) is often costly and cumbersome, mainly due to the synthesis and labeling of isotopes and the requirement of expert knowledge. In this work, we regard molecule identification as a spectrum-to-sequence conversion problem and propose an end-to-end model, called MS$^2$-Transformer, to address this task. The chemical knowledge, defined through a fragmentation tree from the MS/MS spectrum, is incorporated into MS$^2$-Transformer. Our method achieves state-of-the-art results on two widely used benchmarks in molecule identification. To our best knowledge, MS$^2$-Transformer is the first machine learning model that can accurately identify the structures (e.g., molecular graph) from experimental MS/MS rather than chemical formula/categories only (e.g., C$_6$H$_{12}$O$_6$/organic compound), demonstrating it the great application potential in biomedical studies.
****

MetaBalance: High-Performance Neural Networks for Class-Imbalanced Data

Arpit Bansal,Micah Goldblum,Valeriia Cherepanova,Avi Schwarzschild,C. Bayan Bruss,Tom Goldstein

Class-imbalanced data, in which some classes contain far more samples than others, is ubiquitous in real-world applications. Standard techniques for handling class-imbalance usually work by training on a re-weighted loss or on re-balanced data. Unfortunately, training overparameterized neural networks on such objectives causes rapid memorization of minority class data. To avoid this trap, we harness meta-learning, which uses both an "outer-loop'' and an "inner-loop'' loss, each of which may be balanced using different strategies. We evaluate our method, MetaBalance, on image classification, credit-card fraud detection, loan default prediction, and facial recognition tasks with severely imbalanced data. We find that MetaBalance outperforms a wide array of popular strategies designed to handle class-imbalance, especially in scenarios with very few samples in minority classes.
****

Missingness Bias in Model Debugging

Saachi Jain,Hadi Salman,Eric Wong,Pengchuan Zhang,Vibhav Vineet,Sai Vemprala,Aleksander Madry

Missingness, or the absence of features from an input, is a concept fundamental to many model debugging tools. However, in computer vision, pixels cannot simply be removed from an image. One thus tends to resort to heuristics such as blacking out pixels, which may in turn introduce bias into the debugging process. We study such biases and, in particular, show how transformer-based architectures can enable a more natural implementation of missingness, which side-steps these issues and improves the reliability of model debugging in practice.

****

Improved deterministic l2 robustness on CIFAR-10 and CIFAR-100

Sahil Singla,Surbhi Singla,Soheil Feizi

Training convolutional neural networks (CNNs) with a strict Lipschitz constraint under the $l_{2}$ norm is useful for provable adversarial robustness, interpretable gradients and stable training. While $1$-Lipschitz CNNs can be designed by enforcing a $1$-Lipschitz constraint on each layer, training such networks requires each layer to have an orthogonal Jacobian matrix (for all inputs) to prevent the gradients from vanishing during backpropagation. A layer with this property is said to be Gradient Norm Preserving (GNP). In this work, we introduce a procedure to certify the robustness of $1$-Lipschitz CNNs by relaxing the orthogonalization of the last linear layer of the network that significantly advances the state of the art for both standard and provable robust accuracies on CIFAR-100 (gains of $4.80\%$ and $4.71\%$, respectively). We further boost their robustness by introducing (i) a novel Gradient Norm preserving activation function called the Householder activation function (that includes every $\mathrm{GroupSort}$ ac tivation) and (ii) a certificate regularization. On CIFAR-10, we achieve significant improvements over prior works in provable robust accuracy (5.81%) with only a minor drop in standard accuracy (-0.29%). Code for reproducing all experiments in the paper is available at \url{https://github.com/singlasahil14/SOC}.

****

On Adversarial Bias and the Robustness of Fair Machine Learning
Hongyan Chang,Ta Duy Nguyen,Sasi Kumar Murakonda,Ehsan Kazemi,Reza Shokri

Optimizing prediction accuracy can come at the expense of fairness. Towards minimizing discrimination against a group, fair machine learning algorithms strive to equalize the error of a model across different groups, through imposing fairness constraints on the learning algorithm. But, are decisions made by fair models trustworthy? How sensitive are fair models to changes in their training data? By giving equal importance to groups of different sizes and distributions in the training set, we show that fair models become more fragile to outliers. We study the trade-off between fairness and robustness, by analyzing the adversarial (worst-case) bias against group fairness in machine learning and by comparing it with the effect of similar adversarial manipulations on regular models. We show that the adversarial bias introduced in training data, via the sampling or labeling processes, can significantly reduce the test accuracy on fair models, compared with regular models. Our results demonstrate that adversarial bias can also worsen a model's fairness gap on test data, even though the model satisfies the fairness constraint on training data. We analyze the robustness of multiple fair machine learning algorithms that satisfy equalized odds (and equal opportunity) notion of fairness.

****

Beyond Examples: Constructing Explanation Space for Explaining Prototypes
Hyungjun Joo,Seokhyeon Ha,Jae Myung Kim,Sungyeob Han,Jungwoo Lee

As deep learning has been successfully deployed in diverse applications, there is ever increasing need for explaining its decision. Most of the existing methods produced explanations with a second model that explains the first black-box model, but we propose an inherently interpretable model for more faithful explanations. Our method constructs an explanation space in which similarities in terms of human-interpretable features at images share similar latent representations by using a variational autoencoder. This explanation space provides additional explanations of the relationships, going beyond previous classification networks that provide explanations by distances and learned prototypes. In addition, our distance has more intrinsic meaning by VAE training techniques that regulate the latent space. With user study, we validate the quality of explanation space and additional explanations.

****

Learning Global Spatial Information for Multi-View Object-Centric Models
Yuya Kobayashi,Masahiro Suzuki,Yutaka Matsuo

Recently, several studies have been working on multi-view object-centric models, which predict unobserved views of a scene and infer object-centric representations from several observation views. In general, multi-object scenes can be uniquely determined if both the properties of individual objects and the spatial arrangement of objects are specified; however, existing multi-view object-centric models only infer object-level representations and lack spatial information. This insufficient modeling can degrade novel-view synthesis quality and make it difficult to generate novel scenes. We can model both spatial information and object representations by introducing hierarchical probabilistic model, which contains a global latent variable on top of object-level latent variables. However, how to execute inference and training with that hierarchical multi-view object-centric model is unclear. Therefore, we introduce several crucial components which help inference and training with the proposed model. We show that the proposed method achieves good inference quality and can also generate novel scenes.

****

Transition to Linearity of Wide Neural Networks is an Emerging Property of Assembling Weak Models

Chaoyue Liu,Libin Zhu,Misha Belkin

Wide neural networks with linear output layer have been shown to be near-linear, and to have near-constant neural tangent kernel (NTK), in a region containing the optimization path of gradient descent. These findings seem counter-intuitive since in general neural networks are highly complex models. Why does a linear structure emerge when the neural networks become wide?

In this work, we provide a new perspective on this "transition to linearity" by considering a neural network as an assembly model recursively built from a set of sub-models corresponding to individual neurons. In this view, we show that the linearity of wide neural networks is, in fact, an emerging property of assembling a large number of diverse ``weak'' sub-models, none of which dominate the assembly.

****

Coarformer: Transformer for large graph via graph coarsening

Weirui Kuang,Zhen WANG,Yaliang Li,Zhewei Wei,Bolin Ding

Although Transformer has been generalized to graph data, its advantages are mostly observed on small graphs, such as molecular graphs. In this paper, we identify the obstacles of applying Transformer to large graphs: (1) The vast number of distant nodes distract the necessary attention of each target node from its local neighborhood; (2) The quadratic computational complexity regarding the number of nodes makes the learning procedure costly. We get rid of these obstacles by exploiting the complementary natures of GNN and Transformer, and trade the fine-grained long-range information for the efficiency of Transformer. In particular, we present Coarformer, a two-view architecture that captures fine-grained local information using a GNN-based module on the original graph and coarse yet long-range information using a Transformer-based module on the coarse graph (with far fewer nodes). Meanwhile, we design a scheme to enable message passing across these two views to enhance each other. Finally, we conduct extensive experiments on real-world datasets, where Coarformer outperforms any single-view method that solely applies a GNN or Transformer. Besides, the coarse global view and the cross-view propagation scheme enable Coarformer to perform better than the combinations of different GNN-based and Transformer-based modules while consuming the least running time and GPU memory.

****

Delving into Feature Space: Improving Adversarial Robustness by Feature Spectral Regularization

Zhen Cheng,Fei Zhu,Xu-yao Zhang,Cheng-lin Liu

The study of adversarial examples in deep neural networks has attracted great attention. Numerous methods are proposed to eliminate the gap of features between natural examples and adversarial examples. Nevertheless, every feature may play a different role in adversarial robustness. It is worth exploring which feature is more beneficial for robustness. In this paper, we delve into this problem from the perspective of spectral analysis in feature space. We define a new metric to measure the change of features along eigenvectors under adversarial attacks. One key finding is that eigenvectors with smaller eigenvalues are more non-robust, i.e., adversary adds more components along such directions. We attribute this phenomenon to the dominance of the top eigenvalues. To alleviate this problem, we propose a method called \textit{Feature Spectral Regularization (FSR)} to penalize the largest eigenvalue, and as a result, the other smaller eigenvalues get increased relatively. Comprehensive experiments demonstrate that FSR is effective to alleviate the dominance of larger eigenvalues and improve adversarial robustness on different datasets. Our codes will be publicly available soon.

****

Revisiting Virtual Nodes in Graph Neural Networks for Link Prediction

EunJeong Hwang,Veronika Thost,Shib Sankar Dasgupta,Tengfei Ma

It is well known that the graph classification performance of graph neural networks often improves by adding an artificial virtual node to the graphs, which is connected to all nodes in the graph. Intuitively, the virtual node provides a shortcut for message passing between nodes along the graph edges. Surprisingly, the impact of virtual nodes with other problems is still an open research question

.

In this paper, we adapt the concept of virtual nodes to the link prediction scenario, where we usually have much larger, often dense, and more heterogeneous graphs. In particular, we use multiple virtual nodes per graph and graph-based clustering to determine the connections to the graph nodes. We also investigate alternative clustering approaches (e.g., random or more advanced) and compare to the original model with a single virtual node. We conducted extensive experiments over different datasets of the Open Graph Benchmark (OGB) and analyze the results in detail. We show that our virtual node extensions yield rather stable performance increases and allow standard graph neural networks to compete with complex state-of-the-art models, as well as with the models leading the OGB leaderboards.

****

Distributionally Robust Recourse Action

Duy Nguyen,Ngoc Bui,Viet Anh Nguyen

Recourse actions, also known as counterfactual explanations, aim to explain a particular algorithmic decision by showing one or multiple ways in which the instance could be modified to receive an alternate outcome. Existing recourse recommendations often assume that the machine learning models do not change over time. However, this assumption does not always hold in practice because of data distribution shifts, and in this case, the recourse actions may become invalid. To redress this shortcoming, we propose the Distributionally Robust Recourse Action framework, which generates a recourse action that has high probability of being valid under a mixture of model shifts. We show that the robust recourse can be found efficiently using a projected gradient descent algorithm and we discuss several extensions of our framework. Numerical experiments with both synthetic and real-world datasets demonstrate the benefits of our proposed framework.

****

Evolutionary perspective on model fine-tuning

Andrei Kucharavy,Ljiljana Dolamic,Rachid Guerraoui

Be it in natural language generation or in the image generation, massive performances gains have been achieved in the last years. While a substantial part of these advances can be attributed to improvement in machine learning architectures, an important role has also been played by the ever-increasing parameter number of machine learning models, which made from-scratch retraining of the models prohibitively expensive for a large number of users.

In response to that, model fine-tuning - starting with an already good model and further training it on the data relevant to a new, related problem, gained in popularity. This fine-tuning is formally similar to the natural evolution of genetic codes in response to shifting environment.

Here, we formalize this similarity in the framework of Fisher Geometric model and extreme value theory and present a set of tricks used by naturally evolving organisms to accelerate their adaptation, applicable to model fine-tuning.

****

Meta Learning Low Rank Covariance Factors for Energy Based Deterministic Uncertainty

Jeffrey Ryan Willette,Hae Beom Lee,Juho Lee,Sung Ju Hwang

Numerous recent works utilize bi-Lipschitz regularization of neural network layers to preserve relative distances between data instances in the feature spaces of each layer. This distance sensitivity with respect to the data aids in tasks such as uncertainty calibration and out-of-distribution (OOD) detection. In previous works, features extracted with a distance sensitive model are used to construct feature covariance matrices which are used in deterministic uncertainty estimation or OOD detection. However, in cases where there is a distribution over tasks, these methods result in covariances which are sub-optimal, as they may not leverage all of the meta information which can be shared among tasks. With the use of an attentive set encoder, we propose to meta learn either diagonal or diagonal plus low-rank factors to efficiently construct task specific covariance matrices. Additionally, we propose an inference procedure which utilizes scaled ene

rgy to achieve a final predictive distribution which is well calibrated under a distributional dataset shift.
****

Conditional Object-Centric Learning from Video
Thomas Kipf,Gamaleldin Fathy Elsayed,Aravindh Mahendran,Austin Stone,Sara Sabour,Georg Heigold,Rico Jonschkowski,Alexey Dosovitskiy,Klaus Greff

Object-centric representations are a promising path toward more systematic generalization by providing flexible abstractions upon which compositional world models can be built. Recent work on simple 2D and 3D datasets has shown that models with object-centric inductive biases can learn to segment and represent meaningful objects from the statistical structure of the data alone without the need for any supervision. However, such fully-unsupervised methods still fail to scale to diverse realistic data, despite the use of increasingly complex inductive biases such as priors for the size of objects or the 3D geometry of the scene. In this paper, we instead take a weakly-supervised approach and focus on how 1) using the temporal dynamics of video data in the form of optical flow and 2) conditioning the model on simple object location cues can be used to enable segmenting and tracking objects in significantly more realistic synthetic data. We introduce a sequential extension to Slot Attention which we train to predict optical flow for realistic looking synthetic scenes and show that conditioning the initial state of this model on a small set of hints, such as center of mass of objects in the first frame, is sufficient to significantly improve instance segmentation. These benefits generalize beyond the training distribution to novel objects, novel backgrounds, and to longer video sequences. We also find that such initial-state-conditioning can be used during inference as a flexible interface to query the model for specific objects or parts of objects, which could pave the way for a range of weakly-supervised approaches and allow more effective interaction with trained models.
****

An object-centric sensitivity analysis of deep learning based instance segmentation
Johannes Theodoridis,Jessica Hofmann,Johannes Maucher,Andreas Schilling

In this study we establish a comprehensive baseline regarding the object-centric robustness of deep learning models for instance segmentation. Our approach is motivated by the work of Geirhos et al. (2019) on texture bias in CNNs. However, we do not compare against human performance but instead incorporate ideas from object-centric representation learning. In addition, we analyze and control the effect of strong stylization that can lead to disappearing objects. The result is a stylized and object-centric version of MS COCO on which we perform an extensive sensitivity analysis regarding visual feature corruptions. We evaluate a broad range of frameworks including Cascade and Mask R-CNN, Swin Transformer, YOLACT(++), DETR, SOTR and SOLOv2. We find that framework choice, data augmentation and dynamic architectures improve robustness whereas supervised and self supervised pre-training does surprisingly not. In summary we evaluate 63 models on 61 versions of COCO for a total of 3843 evaluations.
****

TAG: Task-based Accumulated Gradients for Lifelong learning
Pranshu Malviya,Balaraman Ravindran,Sarath Chandar

When an agent encounters a continual stream of new tasks in the lifelong learning setting, it leverages the knowledge it gained from the earlier tasks to help learn the new tasks better. In such a scenario, identifying an efficient knowledge representation becomes a challenging problem. Most research works propose to either store a subset of examples from the past tasks in a replay buffer, dedicate a separate set of parameters to each task or penalize excessive updates over parameters by introducing a regularization term. While existing methods employ the general task-agnostic stochastic gradient descent update rule, we propose a task-aware optimizer that adapts the learning rate based on the relatedness among tasks. We utilize the directions taken by the parameters during the updates by additively accumulating the gradients specific to each task. These task-based accumulated gradients act as a knowledge base that is maintained and updated throug hout the stream. We empirically show that our proposed adaptive learning rate not only accounts for catastrophic forgetting but also exhibits knowledge transfer. We also show that our method performs better than several state-of-the-art methods in lifelong learning on complex datasets. Moreover, our method can also be combined with the existing methods and achieve substantial improvement in performance.

****

Gradient Explosion and Representation Shrinkage in Infinite Networks
Adam Klukowski
We study deep fully-connected neural networks using the mean field formalism, and carry out a non-perturbative analysis of signal propagation. As a result, we demonstrate that increasing the depth leads to gradient explosion or to another undesirable phenomenon we call representation shrinkage. The appearance of at least one of these problems is not restricted to a specific initialization scheme or a choice of activation function, but rather is an inherent property of the fully-connected architecture itself. Additionally, we show that many popular normalization techniques fail to mitigate these problems. Our method can also be applied to residual networks to guide the choice of initialization variances.

****

Meta-Learning an Inference Algorithm for Probabilistic Programs
Gwonsoo Che,Hongseok Yang
We present a meta-algorithm for learning a posterior-inference algorithm for restricted probabilistic programs. Our meta-algorithm takes a training set of probabilistic programs that describe models with observations, and attempts to learn an efficient method for inferring the posterior of a similar program.

A key feature of our approach is the use of what we call a white-box inference algorithm that extracts information directly from model descriptions themselves, given as programs. Concretely, our white-box inference algorithm is equipped with multiple neural networks, one for each type of atomic command, and computes an approximate posterior of a given probabilistic program by analysing individual atomic commands in the program using these networks. The parameters of these networks are then learnt from a training set by our meta-algorithm.

We empirically demonstrate that the learnt inference algorithm generalises well to programs that are new in terms of both parameters and structures, and report cases where our approach has advantages over alternative approaches such as HMC in terms of test-time efficiency.

The overall results show the promise as well as remaining challenges of our approach.

****

Calibrated ensembles - a simple way to mitigate ID-OOD accuracy tradeoffs
Ananya Kumar,Aditi Raghunathan,Tengyu Ma,Percy Liang
We often see undesirable tradeoffs in robust machine learning where out-of-distribution (OOD) accuracy is at odds with in-distribution (ID) accuracy. A ‘robust’ classifier obtained via specialized techniques like removing spurious features has better OOD but worse ID accuracy compared to a ‘standard’ classifier trained via vanilla ERM. On six distribution shift datasets, we find that simply ensembling the standard and robust models is a strong baseline---we match the ID accuracy of a standard model with only a small drop in OOD accuracy compared to the robust model. However, calibrating these models in-domain surprisingly improves the OOD accuracy of the ensemble and completely eliminates the tradeoff and we achieve the best of both ID and OOD accuracy over the original models.

****

Variance Reduced Domain Randomization for Policy Gradient
Yuankun Jiang,Chenglin Li,Wenrui Dai,Junni Zou,Hongkai Xiong
By introducing randomness on environment parameters that fundamentally affect the dynamics, domain randomization (DR) imposes diversity to the policy trained by

deep reinforcement learning, and thus improves its capability of generalization. The randomization of environments, however, introduces another source of variability for the estimate of policy gradients, in addition to the already high variance due to trajectory sampling. Therefore, with standard state-dependent baselines, the policy gradient methods may still suffer high variance, causing low sample efficiency during the training of DR. In this paper, we theoretically derive a bias-free and state/environment-dependent optimal baseline for DR, and analytically show its ability to achieve further variance reduction over the standard constant and state-dependent baselines for DR. We further propose a variance reduced domain randomization (VRDR) approach for policy gradient methods, to strike a tradeoff between the variance reduction and computational complexity in practice. By dividing the entire space of environments into some subspaces and estimating the state/subspace-dependent baseline, VRDR enjoys a theoretical guarantee of faster convergence than the state-dependent baseline. We conduct empirical evaluations on six robot control tasks with randomized dynamics. The results demonstrate that VRDR can consistently accelerate the convergence of policy training in all tasks, and achieve even higher rewards in some specific tasks.

****

Scale Efficiently: Insights from Pretraining and Finetuning Transformers

Yi Tay,Mostafa Dehghani,Jinfeng Rao,William Fedus,Samira Abnar,Hyung Won Chung,Sharan Narang,Dani Yogatama,Ashish Vaswani,Donald Metzler

There remain many open questions pertaining to the scaling behaviour of Transformer architectures. These scaling decisions and findings can be critical, as training runs often come with an associated computational cost which have both financial and/or environmental impact. The goal of this paper is to present scaling insights from pretraining and finetuning Transformers. While Kaplan et al. presents a comprehensive study of the scaling behaviour of Transformer language models, the scope is only on the upstream (pretraining) loss. Therefore, it is still unclear if these set of findings transfer to downstream task within the context of the pretrain-finetune paradigm. The key findings of this paper are as follows: (1) we show that aside from only the model size, model shape matters for downstream fine-tuning, (2) scaling protocols operate differently at different compute regions, (3) widely adopted T5-base and T5-large sizes are Pareto-inefficient. To this end, we present improved scaling protocols whereby our redesigned models achieve similar downstream fine-tuning quality while having 50\% fewer parameters and training 40\% faster compared to the widely adopted T5-base model. We publicly release over 100 pretrained checkpoints of different T5 configurations to facilitate future research and analysis.

****

Relative Molecule Self-Attention Transformer

Lukasz Maziarka,Dawid Majchrowski,Tomasz Danel,Piotr Gai■ski,Jacek Tabor,Igor T. Podolak,Pawel Morkisz,Stanislaw Kamil Jastrzebski

Self-supervised learning holds promise to revolutionize molecule property prediction - a central task to drug discovery and many more industries - by enabling data efficient learning from scarce experimental data. Despite significant progress, non-pretrained methods can be still competitive in certain settings. We reason that architecture might be a key bottleneck. In particular, enriching the backbone architecture with domain-specific inductive biases has been key for the success of self-supervised learning in other domains. In this spirit, we methodologically explore the design space of the self-attention mechanism tailored to molecular data. We identify a novel variant of self-attention adapted to processing molecules, inspired by the relative self-attention layer, which involves fusing embedded graph and distance relationships between atoms. Our main contribution is Relative Molecule Attention Transformer (R-MAT): a novel Transformer-based model based on the developed self-attention layer that achieves state-of-the-art or very competitive results across a~wide range of molecule property prediction tasks.

****

Vitruvion: A Generative Model of Parametric CAD Sketches

Ari Seff,Wenda Zhou,Nick Richardson,Ryan P Adams

Parametric computer-aided design (CAD) tools are the predominant way that engineers specify physical structures, from bicycle pedals to airplanes to printed circuit boards. The key characteristic of parametric CAD is that design intent is encoded not only via geometric primitives, but also by parameterized constraints between the elements. This relational specification can be viewed as the construction of a constraint program, allowing edits to coherently propagate to other parts of the design. Machine learning offers the intriguing possibility of accelerating the design process via generative modeling of these structures, enabling new tools such as autocompletion, constraint inference, and conditional synthesis. In this work, we present such an approach to generative modeling of parametric CAD sketches, which constitute the basic computational building blocks of modern mechanical design. Our model, trained on real-world designs from the SketchGraphs dataset, autoregressively synthesizes sketches as sequences of primitives, with initial coordinates, and constraints that reference back to the sampled primitives. As samples from the model match the constraint graph representation used in standard CAD software, they may be directly imported, solved, and edited according to downstream design tasks. In addition, we condition the model on various contexts, including partial sketches (primers) and images of hand-drawn sketches. Evaluation of the proposed approach demonstrates its ability to synthesize realistic CAD sketches and its potential to aid the mechanical design workflow.
****

Bag-of-Vectors Autoencoders for Unsupervised Conditional Text Generation
Florian Mai,James Henderson

Text autoencoders are often used for unsupervised conditional text generation by applying mappings in the latent space to change attributes to the desired values. Recently, Mai et al. (2020) proposed $\operatorname{Emb2Emb}$, a method to $\textit{learn}$ these mappings in the embedding space of an autoencoder. However, their method is restricted to autoencoders with a single-vector embedding, which limits how much information can be retained. We address this issue by extending their method to $\textit{Bag-of-Vectors Autoencoders}$ (BoV-AEs), which encode the text into a variable-size bag of vectors that grows with the size of the text, as in attention-based models. This allows to encode and reconstruct much longer texts than standard autoencoders. Analogous to conventional autoencoders, we propose regularization techniques that facilitate learning meaningful operations in the latent space. Finally, we adapt $\operatorname{Emb2Emb}$ for a training scheme that learns to map an input bag to an output bag, including a novel loss function and neural architecture. Our experimental evaluations on unsupervised sentiment transfer and sentence summarization show that our method performs substantially better than a standard autoencoder.
****

FedGEMS: Federated Learning of Larger Server Models via Selective Knowledge Fusion
Sijie Cheng,Jingwen Wu,Yanghua Xiao,Yang Liu,Yang Liu

Today data is often scattered among billions of resource-constrained edge devices with security and privacy constraints. Federated Learning (FL) has emerged as a viable solution to learn a global model while keeping data private, but the model complexity of FL is impeded by the computation resources of edge nodes. In this work, we investigate a novel paradigm to take advantage of a powerful server model to break through model capacity in FL. By selectively learning from multiple teacher clients and itself, a server model develops in-depth knowledge and transfers its knowledge back to clients in return to boost their respective performance. Our proposed framework achieves superior performance on both server and client models and provides several advantages in a unified framework, including flexibility for heterogeneous client architectures, robustness to poisoning attacks, and communication efficiency between clients and server. By bridging FL effectively with larger server model training, our proposed paradigm paves ways for robust and continual knowledge accumulation from distributed and private data.
****

Space-Time Graph Neural Networks
Samar Hadou,Charilaos I Kanatsoulis,Alejandro Ribeiro

We introduce space-time graph neural network (ST-GNN), a novel GNN architecture, tailored to jointly process the underlying space-time topology of time-varying network data. The cornerstone of our proposed architecture is the composition of time and graph convolutional filters followed by pointwise nonlinear activation functions. We introduce a generic definition of convolution operators that mimic the diffusion process of signals over its underlying support. On top of this definition, we propose space-time graph convolutions that are built upon a composition of time and graph shift operators. We prove that ST-GNNs with multivariate integral Lipschitz filters are stable to small perturbations in the underlying graphs as well as small perturbations in the time domain caused by time warping. Our analysis shows that small variations in the network topology and time evolution of a system does not significantly affect the performance of ST-GNNs. Numerical experiments with decentralized control systems showcase the effectiveness and stability of the proposed ST-GNNs.

****

Scattering Networks on the Sphere for Scalable and Rotationally Equivariant Spherical CNNs

Jason McEwen,Christopher Wallis,Augustine N. Mavor-Parker

Convolutional neural networks (CNNs) constructed natively on the sphere have been developed recently and shown to be highly effective for the analysis of spherical data. While an efficient framework has been formulated, spherical CNNs are nevertheless highly computationally demanding; typically they cannot scale beyond spherical signals of thousands of pixels. We develop scattering networks constructed natively on the sphere that provide a powerful representational space for spherical data. Spherical scattering networks are computationally scalable and exhibit rotational equivariance, while their representational space is invariant to isometries and provides efficient and stable signal representations. By integrating scattering networks as an additional type of layer in the generalized spherical CNN framework, we show how they can be leveraged to scale spherical CNNs to the high-resolution data typical of many practical applications, with spherical signals of many tens of megapixels and beyond.

****

Flow-based Recurrent Belief State Learning for POMDPs

Xiaoyu Chen,Yao Mu,Ping Luo,Shengbo Eben Li,Jianyu Chen

Partially Observable Markov Decision Process (POMDP) provides a principled and generic framework to model real world sequential decision making processes but yet remains unsolved, especially for high dimensional continuous space and unknown models. The main challenge lies in how to accurately obtain the belief state, which is the probability distribution over the unobservable environment states given historical information. Accurately calculating this belief state is a precondition for obtaining an optimal policy of POMDPs. Recent advances in deep learning techniques show great potential to learn good belief states, but they assume the belief states follow certain types of simple distributions such as diagonal Gaussian, which imposes strong restrictions to precisely capture the real belief states. In this paper, we introduce the \textbf{F}l\textbf{O}w-based \textbf{R}ecurrent \textbf{BE}lief \textbf{S}tate model (FORBES), which incorporates normalizing flows into the variational inference to learn general continuous belief states for POMDPs. Furthermore, we show that the learned belief states can be plugged into downstream RL algorithms to improve performance. In experiments, we show that our methods successfully capture the complex belief states that enable multi-modal predictions as well as high quality reconstructions, and results on challenging visual-motor control tasks show that our method achieves superior performance and sample efficiency.

****

Expressiveness and Approximation Properties of Graph Neural Networks

Floris Geerts,Juan L Reutter

Characterizing the separation power of graph neural networks (GNNs) provides an understanding of their limitations for graph learning tasks. Results regarding separation power are, however, usually geared at specific GNNs architectures, and tools for understanding arbitrary GNN architectures are generally lacking. We p rovide an elegant way to easily obtain bounds on the separation power of GNNs in terms of the Weisfeiler-Leman (WL) tests, which have become the yardstick to measure the separation power of GNNs. The crux is to view GNNs as expressions in a procedural tensor language describing the computations in the layers of the GNNs. Then, by a simple analysis of the obtained expressions, in terms of the number of indexes used and the nesting depth of summations, bounds on the separation power in terms of the WL-tests readily follow. We use tensor language to define Higher-Order Message-Passing Neural Networks (or k-MPNNs), a natural extension of MPNNs. Furthermore, the tensor language point of view allows for the derivation of universality results for classes of GNNs in a natural way. Our approach provides a toolbox with which GNN architecture designers can analyze the separation power of their GNNs, without needing to know the intricacies of the WL-tests. We also provide insights in what is needed to boost the separation power of GNNs.

****

Looking Back on Learned Experiences For Class/task Incremental Learning

Mozhgan PourKeshavarzi,Guoying Zhao,Mohammad Sabokrou

Classical deep neural networks are limited in their ability to learn from emerging streams of training data. When trained sequentially on new or evolving tasks, their performance degrades sharply, making them inappropriate in real-world use cases. Existing methods tackle it by either storing old data samples or only updating a parameter set of deep neural networks, which, however, demands a large memory budget or spoils the flexibility of models to learn the incremented task distribution. In this paper, we shed light on an on-call transfer set to provide past experiences whenever a new task arises in the data stream. In particular, we propose a Cost-Free Incremental Learning (CF-IL) not only to replay past experiences the model has learned but also to perform this in a cost free manner. Towards this end, we introduced a memory recovery paradigm in which we query the network to synthesize past exemplars whenever a new task emerges. Thus, our method needs no extra memory for data buffering or network growing, besides calls the proposed memory recovery paradigm to provide past exemplars, named a transfer set in order to mitigate catastrophically forgetting the former tasks in the Incremental Learning (IL) setup. Moreover, in contrast with recently proposed methods, the suggested paradigm does not desire a parallel architecture since it only relies on the learner network. Compared to the state-of-the-art data techniques without buffering past data samples, CF-IL demonstrates significantly better performance on the well-known datasets whether a task oracle is available in test time (Task-IL) or not (Class-IL).

****

Cluster-based Feature Importance Learning for Electronic Health Record Time-series

Henrique Aguiar,Mauro Santos,Peter Watkinson,Tingting Zhu

The recent availability of Electronic Health Records (EHR) has allowed for the development of algorithms predicting inpatient risk of deterioration and trajectory evolution. However, prediction of disease progression with EHR is challenging since these data are sparse, heterogeneous, multi-dimensional, and multi-modal time-series. As such, clustering is used to identify similar groups within the patient cohort to improve prediction. Current models
have shown some success in obtaining cluster representation of patient trajectories, however, they i) fail to obtain clinical interpretability for each cluster, and ii) struggle to learn meaningful cluster numbers in the context of the imbalanced distribution of disease outcomes. We propose a supervised deep learning model to cluster EHR data based on the identification of clinically understandable phenotypes with regard to both outcome prediction and patient trajectory. We introduce novel loss functions to address the problems of class imbalance and cluster collapse, and furthermore propose a feature-time attention mechanism to identify cluster-based phenotype importance across time and feature dimensions. We tested our model in over 100,000 unique trajectories from hospitalised patients with Type-II respiratory failure to predict five different outcomes. Our model yielded added interpretability to cluster formation and outperformed benchmarks by at least 5% in mean AUROC.

****

Marginal Tail-Adaptive Normalizing Flows

Mike Laszkiewicz,Johannes Lederer,Asja Fischer

Learning the tail behavior of a distribution is a notoriously difficult problem. The number of samples from the tail is small, and deep generative models, such as normalizing flows, tend to concentrate on learning the body of the distribution. In this paper, we focus on improving the ability of normalizing flows to correctly capture the tail behavior and, thus, form more accurate models. We prove that the marginal tailedness of a triangular flow can be controlled via the tailedness of the marginals of the base distribution of the normalizing flow. This theoretical insight leads us to a novel type of triangular flows based on learnable base distributions and data-driven permutations. Since the proposed flows preserve marginal tailedness, we call them marginal tail-adaptive flows (mTAFs). An empirical analysis on synthetic data shows that mTAF improves on the robustness and efficiency of vanilla flows and—motivated by our theory—allows to successfully generate tail samples from the distributions. More generally, our experiments affirm that a careful choice of the base distribution is an effective way to introducing inductive biases to normalizing flows.

****

Effective Polynomial Filter Adaptation for Graph Neural Networks

Vijay Lingam,Chanakya Ajit Ekbote,Manan Sharma,Rahul Ragesh,Arun Iyer,SUNDARARAJAN SELLAMANICKAM

Graph Neural Networks (GNNs) exploit signals from node features and the input graph topology to improve node classification task performance. However, these models tend to perform poorly on heterophilic graphs, where connected nodes have different labels. Recently proposed GNNs work across graphs having varying levels of homophily. Among these, models relying on polynomial graph filters have shown promise. We observe that solutions to these polynomial graph filter models are also solutions to an overdetermined system of equations. It suggests that in some instances, the model needs to learn a reasonably high order polynomial. On investigation, we find the proposed models ineffective at learning such polynomials due to their designs. To mitigate this issue, we perform an eigendecomposition of the graph and propose to learn multiple adaptive polynomial filters acting on different subsets of the spectrum. We theoretically and empirically show that our proposed model learns a better filter, thereby improving classification accuracy. We study various aspects of our proposed model including, dependency on the number of eigencomponents utilized, latent polynomial filters learned, and performance of the individual polynomials on the node classification task. We further show that our model is scalable by evaluating over large graphs. Our model achieves performance gains of up to 10% over the state-of-the-art models and outperforms existing polynomial filter-based approaches in general.

****

Bayesian Neural Network Priors Revisited

Vincent Fortuin,Adrià Garriga-Alonso,Sebastian W. Ober,Florian Wenzel,Gunnar Ratsch,Richard E Turner,Mark van der Wilk,Laurence Aitchison

Isotropic Gaussian priors are the de facto standard for modern Bayesian neural network inference. However, it is unclear whether these priors accurately reflect our true beliefs about the weight distributions or give optimal performance. To find better priors, we study summary statistics of neural network weights in networks trained using stochastic gradient descent (SGD). We find that convolutional neural network (CNN) and ResNet weights display strong spatial correlations, while fully connected networks (FCNNs) display heavy-tailed weight distributions. We show that building these observations into priors can lead to improved performance on a variety of image classification datasets. Surprisingly, these priors mitigate the cold posterior effect in FCNNs, but slightly increase the cold posterior effect in ResNets.

****

On the relationship between disentanglement and multi-task learning

Lukasz Maziarka,Aleksandra Nowak,Maciej Wolczyk,Andrzej Bedychaj

One of the main arguments behind studying disentangled representations is the as sumption that they can be easily reused in different tasks. At the same time finding a joint, adaptable representation of data is one of the key challenges in the multi-task learning setting. In this paper, we take a closer look at the relationship between disentanglement and multi-task learning based on hard parameter sharing. We perform a thorough empirical study of the representations obtained by neural networks trained on automatically generated supervised tasks. Using a set of standard metrics we show that disentanglement appears in a natural way during the process of multi-task neural network training.

****

Lightweight Convolutional Neural Networks By Hypercomplex Parameterization

Eleonora Grassucci,Aston Zhang,Danilo Comminiello

Hypercomplex neural networks have proved to reduce the overall number of parameters while ensuring valuable performances by leveraging the properties of Clifford algebras. Recently, hypercomplex linear layers have been further improved by involving efficient parameterized Kronecker products. In this paper, we define the parameterization of hypercomplex convolutional layers to develop lightweight and efficient large-scale convolutional models. Our method grasps the convolution rules and the filters organization directly from data without requiring a rigidly predefined domain structure to follow. The proposed approach is flexible to operate in any user-defined or tuned domain, from 1D to nD regardless of whether the algebra rules are preset. Such a malleability allows processing multidimensional inputs in their natural domain without annexing further dimensions, as done, instead, in quaternion neural networks for 3D inputs like color images. As a result, the proposed method operates with $1/n$ free parameters as regards its analog in the real domain. We demonstrate the versatility of this approach to multiple domains of application by performing experiments on various image as well as audio datasets in which our method outperforms real and quaternion-valued counterparts.

****

Adversarial Rademacher Complexity of Deep Neural Networks

Jiancong Xiao,Yanbo Fan,Ruoyu Sun,Zhi-Quan Luo

Deep neural networks are vulnerable to adversarial attacks. Adversarial training is one of the most effective algorithms to increase the model's robustness. However, the trained models cannot generalize well to the adversarial examples on the test set. In this paper, we study the generalization of adversarial training through the lens of adversarial Rademacher complexity. Current analysis of adversarial Rademacher complexity is up to two-layer neural networks. In adversarial settings, one major difficulty of generalizing these results to deep neural networks is that we cannot peel off the layer as the classical analysis for standard training. We provide a method to overcome this issue and provide upper bounds of adversarial Rademacher complexity of deep neural networks. Similar to the existing bounds of standard Rademacher complexity of neural nets, our bound also includes the product of weight norms. We provide experiments to show that the adversarially trained weight norms are larger than the standard trained weight norms, thus providing an explanation for the bad generalization performance of adversarial training.

****

Bayesian Active Learning with Fully Bayesian Gaussian Processes

Christoffer Riis,Francisco Antunes,Frederik Boe Hüttel,Carlos Lima Azevedo,Francisco C. Pereira

The bias-variance trade-off is a well-known problem in machine learning that only gets more pronounced the less available data there is. When data is scarce, such as in metamodeling, active learning, and Bayesian optimization, neglecting this trade-off can cause inefficient and non-optimal querying, leading to unnecessary data labeling. In this paper, we focus on metamodeling with active learning and the canonical Gaussian Process (GP). We recognize that, for the GP, the bias-variance trade-off regulation 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 a cquisition functions. The first one is a mode-seeking Bayesian variant of Query-by-Committee (B-QBC), and the second is simultaneously mode-seeking and minimizing the predictive variance through a Query by Mixture Gaussian Processes (QB-MGP) formulation. Across seven simulators, we empirically show that B-QBC outperforms the benchmark functions, whereas QB-MGP is the most robust acquisition function and achieves the best accuracy with the fewest iterations. We generally show that incorporating the bias-variance trade-off in the acquisition functions mitigates unnecessary and expensive data labeling.

****

Goal-Directed Planning via Hindsight Experience Replay

Lorenzo Moro,Amarildo Likmeta,Enrico Prati,Marcello Restelli

We consider the problem of goal-directed planning under a deterministic transition model. Monte Carlo Tree Search has shown remarkable performance in solving deterministic control problems. It has been extended from complex continuous domains through function approximators to bias the search of the planning tree in AlphaZero. Nonetheless, these algorithms still struggle with control problems with sparse rewards, such as goal-directed domains, where a positive reward is awarded only when reaching a goal state. In this work, we recast AlphaZero with Hindsight Experience Replay to tackle complex goal-directed planning tasks. We perform a thorough empirical evaluation in several simulated domains, including a novel application to a quantum compiling domain.

****

Effective Uncertainty Estimation with Evidential Models for Open-World Recognition

Charles Corbière,Marc Lafon,Nicolas THOME,Matthieu Cord,Patrick Perez

Reliable uncertainty estimation is crucial when deploying a classifier in the wild. In this paper, we tackle the challenge of jointly quantifying in-distribution and out-of-distribution (OOD) uncertainties. To this end, we leverage the second-order uncertainty representation provided by evidential models and we introduce KLoS, a Kullback–Leibler divergence criterion defined on the class-probability simplex. By keeping the full distributional information, KLoS captures class confusion and lack of evidence in a single score. A crucial property of KLoS is to be a class-wise divergence measure built from in-distribution samples and to not require OOD training data, in contrast to current second-order uncertainty measures. We further design an auxiliary neural network, KLoSNet, to learn a refined criterion directly aligned with the evidential training objective. In the realistic context where no OOD data is available during training, our experiments show that KLoSNet outperforms first-order and second-order uncertainty measures to simultaneously detect misclassifications and OOD samples. When training with OOD samples, we also observe that existing measures are brittle to the choice of the OOD dataset, whereas KLoS remains more robust.

****

Hybrid Random Features

Krzysztof Marcin Choromanski,Han Lin,Haoxian Chen,Arijit Sehanobish,Yuanzhe Ma,Deepali Jain,Jake Varley,Andy Zeng,Michael S Ryoo,Valerii Likhosherstov,Dmitry Kalashnikov,Vikas Sindhwani,Adrian Weller

We propose a new class of random feature methods for linearizing softmax and Gaussian kernels called hybrid random features (HRFs) that automatically adapt the quality of kernel estimation to provide most accurate approximation in the defined regions of interest. Special instantiations of HRFs lead to well-known methods such as trigonometric (Rahimi & Recht, 2007) or (recently introduced in the context of linear-attention Transformers) positive random features (Choromanski et al., 2021). By generalizing Bochner’s Theorem for softmax/Gaussian kernels and leveraging random features for compositional kernels, the HRF-mechanism provides strong theoretical guarantees - unbiased approximation and strictly smaller worst-case relative errors than its counterparts. We conduct exhaustive empirical evaluation of HRF ranging from pointwise kernel estimation experiments, through tests on data admitting clustering structure to benchmarking implicit-attention Transformers (also for downstream Robotics applications), demonstrating its quality in a wide spectrum of machine learning problems.

****

Pretrained Language Model in Continual Learning: A Comparative Study

Tongtong Wu,Massimo Caccia,Zhuang Li,Yuan-Fang Li,Guilin Qi,Gholamreza Haffari

Continual learning (CL) is a setting in which a model learns from a stream of incoming data while avoiding to forget previously learned knowledge. Pre-trained language models (PLMs) have been successfully employed in continual learning of different natural language problems. With the rapid development of many continual learning methods and PLMs, understanding and disentangling their interactions become essential for continued improvement of continual learning performance. In this paper, we thoroughly compare the continual learning performance over the combination of 5 PLMs and 4 CL approaches on 3 benchmarks in 2 typical incremental settings. Our extensive experimental analyses reveal interesting performance differences across PLMs and across CL methods. Furthermore, our representativeness probing analyses dissect PLMs’ performance characteristics in a layer-wise and task-wise manner, uncovering the extent to which their inner layers suffer from forgetting, and the effect of different CL approaches on each layer. Finally, our observations and analyses open up a number of important research questions that will inform and guide the design of effective continual learning techniques.

****

Salient ImageNet: How to discover spurious features in Deep Learning?

Sahil Singla,Soheil Feizi

Deep neural networks can be unreliable in the real world especially when they heavily use {\it spurious} features for their predictions. Focusing on image classifications, we define {\it core features} as the set of visual features that are always a part of the object definition while {\it spurious features} are the ones that are likely to {\it co-occur} with the object but not a part of it (e.g., attribute ``fingers" for class ``band aid"). Traditional methods for discovering spurious features either require extensive human annotations (thus, not scalable), or are useful on specific models. In this work, we introduce a {\it general} framework to discover a subset of spurious and core visual features used in inferences of a general model and localize them on a large number of images with minimal human supervision. Our methodology is based on this key idea: to identify spurious or core \textit{visual features} used in model predictions, we identify spurious or core \textit{neural features} (penultimate layer neurons of a robust model) via limited human supervision (e.g., using top 5 activating images per feature). We then show that these neural feature annotations {\it generalize} extremely well to many more images {\it without} any human supervision. We use the activation maps for these neural features as the soft masks to highlight spurious or core visual features. Using this methodology, we introduce the {\it Salient Imagenet} dataset containing core and spurious masks for a large set of samples from Imagenet. Using this dataset, we show that several popular Imagenet models rely heavily on various spurious features in their predictions, indicating the standard accuracy alone is not sufficient to fully assess model' performance specially in safety-critical applications. Code is available at \url{https://github.com/singlasahil14/salient_imagenet}.

****

Differentiable DAG Sampling

Bertrand Charpentier,Simon Kibler,Stephan Günnemann

We propose a new differentiable probabilistic model over DAGs (DP-DAG). DP-DAG allows fast and differentiable DAG sampling suited to continuous optimization. To this end, DP-DAG samples a DAG by successively (1) sampling a linear ordering of the node and (2) sampling edges consistent with the sampled linear ordering. We further propose VI-DP-DAG, a new method for DAG learning from observational data which combines DP-DAG with variational inference. Hence,VI-DP-DAG approximates the posterior probability over DAG edges given the observed data. VI-DP-DAG is guaranteed to output a valid DAG at any time during training and does not require any complex augmented Lagrangian optimization scheme in contrast to existing differentiable DAG learning approaches. In our extensive experiments, we compare VI-DP-DAG to other differentiable DAG learning baselines on synthetic and real datasets. VI-DP-DAG significantly improves DAG structure and causal mechanism le arning while training faster than competitors.

****

EntQA: Entity Linking as Question Answering

Wenzheng Zhang,Wenyue Hua,Karl Stratos

A conventional approach to entity linking is to first find mentions in a given document and then infer their underlying entities in the knowledge base. A well-known limitation of this approach is that it requires finding mentions without knowing their entities, which is unnatural and difficult. We present a new model that does not suffer from this limitation called $\textbf{EntQA}$, which stands for $\mbox{\textbf{Ent}ity}$ linking as $\mbox{\textbf{Q}uestion}$ $\mbox{\textbf{A}nswering}$. EntQA first proposes candidate entities with a fast retrieval module, and then scrutinizes the document to find mentions of each candidate with a powerful reader module. Our approach combines progress in entity linking with that in open-domain question answering and capitalizes on pretrained models for dense entity retrieval and reading comprehension. Unlike in previous works, we do not rely on a mention-candidates dictionary or large-scale weak supervision. EntQA achieves strong results on the GERBIL benchmarking platform.

****

Evaluating Model-Based Planning and Planner Amortization for Continuous Control

Arunkumar Byravan,Leonard Hasenclever,Piotr Trochim,Mehdi Mirza,Alessandro Davide Ialongo,Yuval Tassa,Jost Tobias Springenberg,Abbas Abdolmaleki,Nicolas Heess,Josh Merel,Martin Riedmiller

There is a widespread intuition that model-based control methods should be able to surpass the data efficiency of model-free approaches. In this paper we attempt to evaluate this intuition on various challenging locomotion tasks. We take a hybrid approach, combining model predictive control (MPC) with a learned model and model-free policy learning; the learned policy serves as a proposal for MPC. We show that MPC with learned proposals and models (trained on the fly or transferred from related tasks) can significantly improve performance and data efficiency with respect to model-free methods. However, we find that well-tuned model-free agents are strong baselines even for high DoF control problems. Finally, we show that it is possible to distil a model-based planner into a policy that amortizes the planning computation without any loss of performance.

****

Loss meta-learning for forecasting

Alan Collet,Antonio Bazco-Nogueras,Albert Banchs,Marco Fiore

Meta-learning of loss functions for supervised learning has been used to date for classification tasks, or as a way to enable few-shot learning. In this paper, we show how a fairly simple loss meta-learning approach can substantially improve regression results. Specifically, we target forecasting of time series and explore case studies grounded on real-world data, and show that meta-learned losses can benefit the quality of the prediction both in cases that are apparently naive and in practical scenarios where the performance metric is complex, time-correlated, non-differentiable, or not known a-priori.

****

Towards Understanding Data Values: Empirical Results on Synthetic Data

Danilo Brajovic,Omar De Mitri,Alex Windberger,Marco Huber

Understanding the influence of data on machine learning models is an emerging research field. Inspired by recent work in data valuation, we perform several experiments to get an intuition for this influence on a multi-layer perceptron. We generate a synthetic two-dimensional data set to visualize how different valuation methods value data points on a mesh grid spanning the relevant feature space. In this setting, individual data values can be derived directly from the impact of the respective data points on the decision boundary. Our results show that the most important data points are the miss-classified ones. Furthermore, despite performance differences on real world data sets, all investigated methods except one qualitatively agree on the data values derived from our experiments. Finall y, we place our results into the recent literature and discuss data values and their relationship to other methods.

****

Hierarchical Few-Shot Imitation with Skill Transition Models

Kourosh Hakhamaneshi,Ruihan Zhao,Albert Zhan,Pieter Abbeel,Michael Laskin

A desirable property of autonomous agents is the ability to both solve long-horizon problems and generalize to unseen tasks. Recent advances in data-driven skill learning have shown that extracting behavioral priors from offline data can enable agents to solve challenging long-horizon tasks with reinforcement learning. However, generalization to tasks unseen during behavioral prior training remains an outstanding challenge. To this end, we present Few-shot Imitation with Skill Transition Models (FIST), an algorithm that extracts skills from offline data and utilizes them to generalize to unseen tasks given a few downstream demonstrations. FIST learns an inverse skill dynamics model, a distance function, and utilizes a semi-parametric approach for imitation. We show that FIST is capable of generalizing to new tasks and substantially outperforms prior baselines in navigation experiments requiring traversing unseen parts of a large maze and 7-DoF robotic arm experiments requiring manipulating previously unseen objects in a kitchen.

****

Bolstering Stochastic Gradient Descent with Model Building

Ilker Birbil,Özgür Martin,Gönenc Onay,Figen Öztoprak

Stochastic gradient descent method and its variants constitute the core optimization algorithms that achieve good convergence rates for solving machine learning problems. These rates are obtained especially when these algorithms are fine-tuned for the application at hand. Although this tuning process can require large computational costs, recent work has shown that these costs can be reduced by line search methods that iteratively adjust the stepsize. We propose an alternative approach to stochastic line search by using a new algorithm based on forward step model building. This model building step incorporates a second-order information that allows adjusting not only the stepsize but also the search direction. Noting that deep learning model parameters come in groups (layers of tensors), our method builds its model and calculates a new step for each parameter group. This novel diagonalization approach makes the selected step lengths adaptive. We provide convergence rate analysis, and experimentally show that the proposed algorithm achieves faster convergence and better generalization in most problems. Moreover, our experiments show that the proposed method is quite robust as it converges for a wide range of initial stepsizes.

****

Dense-to-Sparse Gate for Mixture-of-Experts

Xiaonan Nie,Shijie Cao,Xupeng Miao,Lingxiao Ma,Jilong Xue,Youshan Miao,Zichao Yang,Zhi Yang,Bin CUI

Mixture-of-experts (MoE) is becoming popular due to its success in improving the model quality, especially in Transformers. By routing tokens with a sparse gate to a few experts that each only contains part of the full model, MoE keeps the model size unchanged and significantly reduces per-token computation, which effectively scales neural networks. However, we found that the current approach of jointly training experts and the sparse gate introduces a negative impact on model accuracy, diminishing the efficiency of expensive large-scale model training. In this work, we proposed $\texttt{Dense-To-Sparse}$ gate (DTS-Gate) for MoE training. Specifically, instead of using a permanent sparse gate, DTS-Gate begins as a dense gate that routes tokens to all experts, then gradually and adaptively becomes sparser while routes to fewer experts. MoE with DTS-Gate naturally decouples the training of experts and the sparse gate by training all experts at first and then learning the sparse gate. Our code is available at https://anonymous.4open.science/r/MoE-3D0D/README.md/README.moe.md.

****

Thinking Deeper With Recurrent Networks: Logical Extrapolation Without Overthinking

Arpit Bansal,Avi Schwarzschild,Eitan Borgnia,Zeyad Emam,Furong Huang,Micah Goldb lum,Tom Goldstein

Classical machine learning systems perform best when they are trained and tested on the same distribution, and they lack a mechanism to increase model power after training is complete. In contrast, recent work has observed that recurrent networks can exhibit logical extrapolation; models trained only on small/simple problem instances can extend their abilities to solve large/complex instances at test time simply by performing more recurrent iterations. While preliminary results on these ``thinking systems'' are promising, existing recurrent systems, when iterated many times, often collapse rather than improve their performance.  This ``overthinking'' phenomenon has prevented thinking systems from scaling to particularly large and complex problems. In this paper, we design a recall architecture that keeps an explicit copy of the problem instance in memory so that it cannot be forgotten. We also propose an incremental 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. Together, these design choices encourage models to converge to a steady state solution rather than deteriorate when many iterations are used. These innovations help to tackle the overthinking problem and boost deep thinking behavior on each of the benchmark tasks proposed by Schwarzschild et al. (2021a).

****

End-to-End Learning of Probabilistic Hierarchies on Graphs

Daniel Zügner,Bertrand Charpentier,Morgane Ayle,Sascha Geringer,Stephan Günnemann

We propose a novel probabilistic model over hierarchies on graphs obtained by continuous relaxation of tree-based hierarchies. We draw connections to Markov chain theory, enabling us to perform hierarchical clustering by efficient end-to-end optimization of relaxed versions of quality metrics such as Dasgupta cost or Tree-Sampling Divergence (TSD).

We show that our model learns rich, high-quality hierarchies present in 11 real world graphs, including a large graph with 2.3M nodes. Our model consistently outperforms recent as well as strong traditional baselines such as average linkage.

Our model also obtains strong results on link prediction despite not being trained on this task, highlighting the quality of the hierarchies discovered by our model.

****

Tessellated 2D Convolution Networks: A Robust Defence against Adversarial Attacks

Swarnava Das,Pabitra Mitra,Debasis Ganguly

Data-driven (deep) learning approaches for image classification are prone to adversarial attacks. This means that an adversarial crafted image which is sufficiently close (visually indistinguishable) to its representative class can often be misclassified to be a member of a different class. A reason why deep neural approaches exhibits such vulnerability towards adversarial threats is mainly because the abstract representations learned in a data-driven manner often do not correlate well with human perceived features. To mitigate this problem, we propose the tessellated 2d convolution network, a novel divide-and-conquer based approach, which first independently learns the abstract representations of non-overlapping regions within an image, and then learns how to combine these representations to infer its class. It turns out that a non-uniform tiling of an image which ensures that the difference between the maximum and the minimum region sizes is not too large is the most robust way to construct such a tessellated 2d convolution network. This criterion can be achieved, among other schemes, by using a Mondrian tessellation of the input image. Our experiments demonstrate that our tessellated networks provides a more robust defence mechanism against gradient-based adversarial attacks in comparison to conventional deep neural models.

****

GeneDisco: A Benchmark for Experimental Design in Drug Discovery

Arash Mehrjou,Ashkan Soleymani,Andrew Jesson,Pascal Notin,Yarin Gal,Stefan Bauer,Patrick Schwab

In vitro cellular experimentation with genetic interventions, using for example CRISPR technologies, is an essential step in early-stage drug discovery and target validation that serves to assess initial hypotheses about causal associations between biological mechanisms and disease pathologies. With billions of potential hypotheses to test, the experimental design space for in vitro genetic experiments is extremely vast, and the available experimental capacity - even at the largest research institutions in the world - pales in relation to the size of this biological hypothesis space. Machine learning methods, such as active and reinforcement learning, could aid in optimally exploring the vast biological space by integrating prior knowledge from various information sources as well as extrapolating to yet unexplored areas of the experimental design space based on available data. However, there exist no standardised benchmarks and data sets for this challenging task and little research has been conducted in this area to date. Here, we introduce GeneDisco, a benchmark suite for evaluating active learning algorithms for experimental design in drug discovery. GeneDisco contains a curated set of multiple publicly available experimental data sets as well as open-source implementations of state-of-the-art active learning policies for experimental design and exploration.
****

GraphENS: Neighbor-Aware Ego Network Synthesis for Class-Imbalanced Node Classification

Joonhyung Park,Jaeyun Song,Eunho Yang

In many real-world node classification scenarios, nodes are highly class-imbalanced, where graph neural networks (GNNs) can be readily biased to major class instances. Albeit existing class imbalance approaches in other domains can alleviate this issue to some extent, they do not consider the impact of message passing between nodes. In this paper, we hypothesize that overfitting to the neighbor sets of minor class due to message passing is a major challenge for class-imbalanced node classification. To tackle this issue, we propose GraphENS, a novel augmentation method that synthesizes the whole ego network for minor class (minor node and its one-hop neighbors) by combining two different ego networks based on their similarity. Additionally, we introduce a saliency-based node mixing method to exploit the abundant class-generic attributes of other nodes while blocking the injection of class-specific features. Our approach consistently outperforms the baselines over multiple node classification benchmark datasets and architectures.
****

Manifold Distance Judge, an Adversarial Samples Defense Strategy Based on Service Orchestration

Mengxin Zhang,Xiaofeng QIU

Deep neural networks (DNNs) are playing an increasingly significant role in the modern world. However, they are weak to adversarial examples that are generated by adding specially crafted perturbations. Most defenses against adversarial examples focused on refining the DNN models, which often sacrifice the performance and computational cost of models on benign samples. In this paper, we propose a manifold distance detection method to distinguish between legitimate samples and adversarial samples by measuring the different distances on the manifold. The manifold distance detection method neither modifies the protected models nor requires knowledge of the process for generating adversarial samples. Inspired by the effectiveness of the manifold distance detection, we demonstrated a well-designed orchestrated defense strategy, named Manifold Distance Judge (MDJ), which selects the best image processing method that will effectively expand the manifold distance between legitimate and adversarial samples, and thus, enhances the performance of the following manifold distance detection method. Tests on the ImageNet dataset, the MDJ is effective against the most adversarial samples under whitebox, graybox, and blackbox attack scenarios. We show empirically that the orchestration strategy MDJ is significantly better than Feature Squeezing on the recall rate. Meanwhile, MDJ achieves high detection rates against CW attack and DI-FGSM attack.
****

Understanding Graph Learning with Local Intrinsic Dimensionality
Xiaojun Guo,Xingjun Ma,Yisen Wang
Many real-world problems can be formulated as graphs and solved by graph learning techniques. Whilst the rise of Graph Neural Networks (GNNs) has greatly advanced graph learning, there is still a lack of understanding of the intrinsic properties of graph data and their impact on graph learning. In this paper, we narrow the gap by studying the intrinsic dimension of graphs with \emph{Local Intrinsic Dimensionality (LID)}. The LID of a graph measures the expansion rate of the graph as the local neighborhood size of the nodes grows.
With LID, we estimate and analyze the intrinsic dimensions of node features, graph structure and representations learned by GNNs. We first show that feature LID (FLID) and structure LID (SLID) are well correlated with the complexity of synthetic graphs. Following this, we conduct a comprehensive analysis of 12 popular graph datasets of diverse categories and show that 1) graphs of lower FLIDs and SLIDs are generally easier to learn; 2) GNNs learn by mapping graphs (feature and structure together) to low-dimensional manifolds that are of much lower representation LIDs (RLIDs), i.e., RLID $\ll$ FLID/SLID; and 3) when the layers go deep in message-passing based GNNs, the underlying graph will converge to a complete graph of $\operatorname{SLID}=0.5$, losing structural information and causing the over-smoothing problem.
****
Continuously Discovering Novel Strategies via Reward-Switching Policy Optimization
Zihan Zhou,Wei Fu,Bingliang Zhang,Yi Wu
We present Reward-Switching Policy Optimization (RSPO), a paradigm to discover diverse strategies in complex RL environments by iteratively finding novel policies that are both locally optimal and sufficiently different from existing ones. To encourage the learning policy to consistently converge towards a previously undiscovered local optimum, RSPO switches between extrinsic and intrinsic rewards via a trajectory-based novelty measurement during the optimization process. When a sampled trajectory is sufficiently distinct, RSPO performs standard policy optimization with extrinsic rewards. For trajectories with high likelihood under existing policies, RSPO utilizes an intrinsic diversity reward to promote exploration. Experiments show that RSPO is able to discover a wide spectrum of strategies in a variety of domains, ranging from single-agent navigation tasks and MuJoCo control to multi-agent stag-hunt games and the StarCraft II Multi-Agent Challenge.
****
Langevin Autoencoders for Learning Deep Latent Variable Models
Shohei Taniguchi,Yusuke Iwasawa,Wataru Kumagai,Yutaka Matsuo
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 since it is not scalable to data size owing to its datapoint-wise iterations and slow convergence. This paper proposes the amortized Langevin dynamics (ALD), wherein datapoint-wise MCMC iterations are entirely replaced with updates of an inference model that maps observations into latent variables. Since it no longer depends on datapoint-wise iterations, ALD enables scalable inference from large-scale datasets. Despite its efficiency, it retains the excellent property of MCMC; we prove that ALD has the target posterior as a stationary distribution with a mild assumption. Furthermore, ALD can be extended to sampling from an unconditional distribution such as an energy-based model, enabling more flexible generative modeling by applying it to the prior distribution of the latent variable. Based on ALD, we construct a new deep latent variable model named the Langevin autoencoder (LAE). LAE uses ALD for autoencoder-like posterior inference and sampling from the latent space EBM. Using toy datasets, we empirically validate that ALD can properly obtain samples from target distributions in both conditional and unconditional cases, and ALD converges significantly faster than traditional LD. We also evaluate LAE on the image generation task using three datasets (SVHN, CIFAR-10, and CelebA-HQ). Not only can LAE be trained faster than non-amortized MCMC methods, but LAE can also generate better samp

les in terms of the Fréchet Inception Distance (FID) compared to AVI-based methods, such as the variational autoencoder.

****

Filtered-CoPhy: Unsupervised Learning of Counterfactual Physics in Pixel Space
Steeven JANNY,Fabien Baradel,Natalia Neverova,Madiha Nadri,Greg Mori,Christian Wolf

Learning causal relationships in high-dimensional data (images, videos) is a hard task, as they are often defined on low dimensional manifolds and must be extracted from complex signals dominated by appearance, lighting, textures and also spurious correlations in the data. We present a method for learning counterfactual reasoning of physical processes in pixel space, which requires the prediction of the impact of interventions on initial conditions. Going beyond the identification of structural relationships, we deal with the challenging problem of forecasting raw video over long horizons. Our method does not require the knowledge or supervision of any ground truth positions or other object or scene properties. Our model learns and acts on a suitable hybrid latent representation based on a combination of dense features, sets of 2D keypoints and an additional latent vector per keypoint. We show that this better captures the dynamics of physical processes than purely dense or sparse representations. We introduce a new challenging and carefully designed counterfactual benchmark for predictions in pixel space and outperform strong baselines in physics-inspired ML and video prediction.

****

Sampling from Discrete Energy-Based Models with Quality/Efficiency Trade-offs
Bryan Eikema,Germán Kruszewski,Hady Elsahar,Marc Dymetman

Energy-Based Models (EBMs) allow for extremely flexible specifications of probability distributions. However, they do not provide a mechanism for obtaining exact samples from these distributions. Monte Carlo techniques can aid us in obtaining samples if some proposal distribution that we can easily sample from is available. For instance, rejection sampling can provide exact samples but is often difficult or impossible to apply due to the need to find a proposal distribution that upper-bounds the target distribution everywhere. Approximate Markov chain Monte Carlo sampling techniques like Metropolis-Hastings are usually easier to design, exploiting a local proposal distribution that performs local edits on an evolving sample. However, these techniques can be inefficient due to the local nature of the proposal distribution and do not provide an estimate of the quality of their samples. In this work, we propose a new approximate sampling technique, Quasi Rejection Sampling (QRS), that allows for a trade-off between sampling efficiency and sampling quality, while providing explicit convergence bounds and diagnostics. QRS capitalizes on the availability of high-quality global proposal distributions obtained from deep learning models. We demonstrate the effectiveness of QRS sampling for discrete EBMs over text for the tasks of controlled text generation with distributional constraints and paraphrase generation. We show that we can sample from such EBMs with arbitrary precision at the cost of sampling efficiency.

****

Lottery Image Prior
Qiming Wu,Xiaohan Chen,Yifan Jiang,Pan Zhou,Zhangyang Wang

Deep Neural Networks (DNNs), either pre-trained (e.g., GAN generator) or untrained (e.g., deep image prior), could act as overparameterized image priors that help solve various image inverse problems. Since traditional image priors have much fewer parameters, those DNN-based priors naturally invite the curious question: do they really have to be heavily parameterized? Drawing inspirations from the recently prosperous research on lottery ticket hypothesis (LTH), we conjecture and study a novel “lottery image prior” (LIP), stated as: given an (untrained or trained) DNN-based image prior, it will have a sparse subnetwork that can be training in isolation, to match the original DNN’s performance when being applied as a prior to various image inverse problems. We conduct extensive experiments in two representative settings: (i) image restoration with the deep image prior, using an untrained DNN; and (ii) compressive sensing image reconstruction, using a pre-trained GAN generator. Our results validate the prevailing existence of L

IP, and that it can be found by iterative magnitude pruning (IMP) with surrogate tasks. Specifically, we can successfully locate the LIP subnetworks at the sparsity range of 20%-86.58% in setting i; and those at sparsity range of 5%-36% in setting ii. Those LIP subnetworks also possess high transferrability. To our best knowledge, this is the first time that LTH is demonstrated to be relevant in the context of inverse problems or image priors, and such compact DNN-based priors may potentially contribute to practical efficiency. Code will be publicly available.

****

Improving Neural Network Generalization via Promoting Within-Layer Diversity

Firas Laakom,Jenni Raitoharju,Alexandros Iosifidis,Moncef Gabbouj

Neural networks are composed of multiple layers arranged in a hierarchical structure jointly trained with a gradient-based optimization, where the errors are back-propagated from the last layer back to the first one. At each optimization step, neurons at a given layer receive feedback from neurons belonging to higher layers of the hierarchy. In this paper, we propose to complement this traditional 'between-layer' feedback with additional 'within-layer' feedback to encourage the diversity of the activations within the same layer. To this end, we measure the pairwise similarity between the outputs of the neurons and use it to model the layer's overall diversity. By penalizing similarities and promoting diversity, we encourage each unit within the layer to learn a distinctive representation and, thus, to enrich the data representation learned and to increase the total capacity of the model. We theoretically study how the within-layer activation diversity affects the generalization performance of a neural network and prove that increasing the diversity of hidden activations reduces the estimation error. In addition to the theoretical guarantees, we present an extensive empirical study confirming that the proposed approach enhances the performance of state-of-the-art neural network models and decreases the generalization gap in multiple tasks.

****

Learning to Remember Patterns: Pattern Matching Memory Networks for Traffic Forecasting

Hyunwook Lee,Seungmin Jin,Hyeshin Chu,Hongkyu Lim,Sungahn Ko

Traffic forecasting is a challenging problem due to complex road networks and sudden speed changes caused by various events on roads. Several models have been proposed to solve this challenging problem, with a focus on learning the spatio-temporal dependencies of roads. In this work, we propose a new perspective for converting the forecasting problem into a pattern-matching task, assuming that large traffic data can be represented by a set of patterns. To evaluate the validity of this new perspective, we design a novel traffic forecasting model called Pattern-Matching Memory Networks (PM-MemNet), which learns to match input data to representative patterns with a key-value memory structure. We first extract and cluster representative traffic patterns that serve as keys in the memory. Then, by matching the extracted keys and inputs, PM-MemNet acquires the necessary information on existing traffic patterns from the memory and uses it for forecasting. To model the spatio-temporal correlation of traffic, we proposed a novel memory architecture, GCMem, which integrates attention and graph convolution. The experimental results indicate that PM-MemNet is more accurate than state-of-the-art models, such as Graph WaveNet, with higher responsiveness. We also present a qualitative analysis describing how PM-MemNet works and achieves higher accuracy when road speed changes rapidly.

****

Federated Learning with GAN-based Data Synthesis for Non-IID Clients

Zijian Li,Jiawei Shao,Yuyi Mao,Jessie Hui Wang,Jun Zhang

Federated learning (FL) has recently emerged as a popular privacy-preserving collaborative learning paradigm. However, it suffers from the non-IID (independent and identically distributed) data among clients. In this paper, we propose a novel framework, namely Synthetic Data Aided Federated Learning (SDA-FL), to resolve the non-IID issue by sharing differentially private synthetic data. Specifically, each client pretrains a local generative adversarial network (GAN) to genera te synthetic data, which are uploaded to the parameter server (PS) to construct a global shared synthetic dataset. The PS is responsible for generating and updating high-quality labels for the global dataset via pseudo labeling with a confident threshold before each global aggregation. A combination of the local private dataset and labeled synthetic dataset leads to nearly identical data distributions among clients, which improves the consistency among local models and benefits the global aggregation. To ensure privacy, the local GANs are trained with differential privacy by adding artificial noise to the local model gradients before being uploaded to the PS. Extensive experiments evidence that the proposed framework outperforms the baseline methods by a large margin in several benchmark datasets under both the supervised and semi-supervised settings.

****

Learning an Ethical Module for Bias Mitigation of pre-trained Models
Jean-Rémy Conti,Nathan Noiry,Stephan CLEMENCON,Vincent Despiegel,Stéphane Gentric

In spite of the high performance and reliability of deep learning algorithms in broad range everyday applications, many investigations tend to show that a lot of models exhibit biases, discriminating against some subgroups of the population. This urges the practitioner to develop fair systems whose performances are uniform among individuals. In this work, we propose a post-processing method designed to mitigate bias of state-of-the-art models. It consists in learning a shallow neural network, called the Ethical Module, which transforms the deep embeddings of a pre-trained model in order to give more representation power to the disadvantaged subgroups. Its training is supervised by the von Mises-Fisher loss, whose hyperparameters allow to control the space allocated to each subgroup in the latent space. Besides being very simple, the resulting methodology is more stable and faster than most current bias mitigation methods. In order to illustrate our idea in a concrete use case, we focus here on gender bias in facial recognition and conduct extensive numerical experiments on standard datasets.

****

Why Propagate Alone? Parallel Use of Labels and Features on Graphs
Yangkun Wang,Jiarui Jin,Weinan Zhang,Yang Yongyi,Jiuhai Chen,Quan Gan,Yong Yu,Zheng Zhang,Zengfeng Huang,David Wipf

One of the challenges of graph-based semi-supervised learning over ordinary supervised learning for classification tasks lies in label utilization. The direct use of ground-truth labels in graphs for training purposes can result in a parametric model learning trivial degenerate solutions (e.g., an identity mapping from input to output). In addressing this issue, a label trick has recently been proposed in the literature and applied to a wide range of graph neural network (GNN) architectures, achieving state-of-the-art results on various datasets. The essential idea is to randomly split the observed labels on the graph and use a fraction of them as input to the model (along with original node features), and predict the remaining fraction. Despite its success in enabling GNNs to propagate features and labels simultaneously, this approach has never been analyzed from a theoretical perspective, nor fully explored across certain natural use cases. In this paper, we demonstrate that under suitable settings, this stochastic trick can be reduced to a more interpretable deterministic form, allowing us to better explain its behavior, including an emergent regularization effect, and motivate broader application scenarios. Our experimental results corroborate these analyses while also demonstrating improved node classification performance applying the label trick in new domains.

****

Learning by Directional Gradient Descent
David Silver,Anirudh Goyal,Ivo Danihelka,Matteo Hessel,Hado van Hasselt

How should state be constructed from a sequence of observations, so as to best achieve some objective? Most deep learning methods update the parameters of the state representation by gradient descent. However, no prior method for computing the gradient is fully satisfactory, for example consuming too much memory, introducing too much variance, or adding too much bias. In this work, we propose a new learning algorithm that addresses these limitations. The basic idea is to upda te the parameters of the representation by using the directional derivative along a candidate direction, a quantity that may be computed online with the same computational cost as the representation itself. We consider several different choices of candidate direction, including random selection and approximations to the true gradient, and investigate their performance on several synthetic tasks.

****

Stepping Back to SMILES Transformers for Fast Molecular Representation Inference
Wenhao Zhu,Ziyao Li,Lingsheng Cai,Guojie Song

In the intersection of molecular science and deep learning, tasks like virtual screening have driven the need for a high-throughput molecular representation generator on large chemical databases. However, as SMILES strings are the most common storage format for molecules, using deep graph models to extract molecular feature from raw SMILES data requires an SMILES-to-graph conversion, which significantly decelerates the whole process. Directly deriving molecular representations from SMILES is feasible, yet there exists a large performance gap between the existing SMILES-based models and graph-based models at benchmark results. To address this issue, we propose ST-KD, an end-to-end SMILES Transformer for molecular representation learning boosted by Knowledge Distillation. In order to conduct knowledge transfer from graph Transformers to ST-KD, we have redesigned the attention layers and introduced a pre-transformation step to tokenize the SMILES strings and inject structure-based positional embeddings. ST-KD shows competitive results on latest standard molecular datasets PCQM4M-LSC and QM9, with $3\text{-}14\times$ inference speed compared with existing graph models.

****

Interventional Black-Box Explanations
Ola Ahmad,Simon Corbeil,Vahid Hashemi,Freddy Lecue

Deep Neural Networks (DNNs) are powerful systems able to freely evolve on their own from training data. However, like any highly parametrized mathematical model, capturing the explanation of any prediction of such models is rather difficult. We believe that there exist relevant mechanisms inside the structure of post-hoc DNNs that supports transparency and interpretability. To capture these mechanisms, we quantify the effects of parameters (pieces of knowledge) on models' predictions using the framework of causality. We introduce a general formalism of the causal diagram to express cause-effect relations inside the DNN's architecture. Then, we develop a novel algorithm to construct explanations of DNN's predictions using the do-operator. We call our method, Interventional Black-Box Explanations. On image classification tasks, we explain the behaviour of the model and extract visual explanations from the effects of the causal filters in convolution layers. We qualitatively demonstrate that our method captures more informative concepts compared to traditional attribution-based methods.

Finally, we believe that our method is orthogonal to logic-based explanation methods and can be leveraged to improve their explanations.

****

Maximum Entropy RL (Provably) Solves Some Robust RL Problems
Benjamin Eysenbach,Sergey Levine

Many potential applications of reinforcement learning (RL) require guarantees that the agent will perform well in the face of disturbances to the dynamics or reward function. In this paper, we prove theoretically that maximum entropy (MaxEnt) RL maximizes a lower bound on a robust RL objective, and thus can be used to learn policies that are robust to some disturbances in the dynamics and the reward function. While this capability of MaxEnt RL has been observed empirically in prior work, to the best of our knowledge our work provides the first rigorous proof and theoretical characterization of the MaxEnt RL robust set. While a number of prior robust RL algorithms have been designed to handle similar disturbances to the reward function or dynamics, these methods typically require additional moving parts and hyperparameters on top of a base RL algorithm. In contrast, our results suggest that MaxEnt RL by itself is robust to certain disturbances, without requiring any additional modifications. While this does not imply that MaxEnt RL is the best available robust RL method, MaxEnt RL is a simple robust RL m

ethod with appealing formal guarantees.
****
WaveCorr: Deep Reinforcement Learning with Permutation Invariant Policy Networks for Portfolio Management
Saeed Marzban,Erick Delage,Jonathan Li
The problem of portfolio management represents an important and challenging class of dynamic decision making problems, where rebalancing decisions need to be made over time with the consideration of many factors such as investors’ preferences, trading environment, and market conditions. In this paper, we present a new portfolio policy network architecture for deep reinforcement learning (DRL) that can exploit more effectively cross-asset dependency information and achieve better performance than state-of-the-art architectures. In doing so, we introduce a new form of permutation invariance property for policy networks and derive general theory for verifying its applicability. Our portfolio policy network, named WaveCorr, is the first convolutional neural network architecture that preserves this invariance property when treating asset correlation information. Finally, in a set of experiments conducted using data from both Canadian (TSX) and American stock markets (S\&P 500), WaveCorr consistently outperforms other architectures with an impressive 3\%-25\% absolute improvement in terms of average annual return, and up to more than 200\% relative improvement in average Sharpe ratio. We also measured an improvement of a factor of up to 5 in the stability of performance under random choices of initial asset ordering and weights. The stability of the network has been found as particularly valuable by our industrial partner.
****
Determining the Ethno-nationality of Writers Using Written English Text
Deenuka Niroshini Perera,Ruvan Weerasinghe,Randhil Pushpananda
Ethno-nationality is where nations are defined by a shared heritage, for instance it can be a membership of a common language, nationality, religion or an ethnic ancestry. The main goal of this research is to determine a person’s country-of-origin using English text written in less controlled environments, employing Machine Learning (ML) and Natural Language Processing (NLP) techniques. The current literature mainly focuses on determining the native language of English writers and a minimal number of researches have been conducted in determining the country-of-origin of English writers.

Further, most experiments in the literature are mainly based on the TOEFL, ICLE datasets which were collected in more controlled environments (i.e., standard exam answers). Hence, most of the writers try to follow some guidelines and patterns of writing. Subsequently, the creativity, freedom of writing and the insights of writers could be hidden. Thus, we believe it hides the real nativism of the writers. Further, those corpora are not freely available as it involves a high cost of licenses. Thus, the main data corpus used for this research was the International Corpus of English (ICE corpus). Up to this point, none of the researchers have utilised the ICE corpus for the purpose of determining the writers’ country-of-origin, even though there is a true potential.

For this research, an overall accuracy of 0.7636 for the flat classification (for all ten countries) and accuracy of 0.6224~1.000 for sub-categories were received. In addition, the best ML model obtained for the flat classification strategy is linear SVM with SGD optimizer trained with word (1,1) uni-gram model.
****
A Unified Contrastive Energy-based Model for Understanding the Generative Ability of Adversarial Training
Yifei Wang,Yisen Wang,Jiansheng Yang,Zhouchen Lin
Adversarial Training (AT) is known as an effective approach to enhance the robustness of deep neural networks. Recently researchers notice that robust models with AT have good generative ability and can synthesize realistic images, while the reason behind it is yet under-explored. In this paper, we demystify this phenomenon by developing a unified probabilistic framework, called Contrastive Energy

-based Models (CEM). On the one hand, we provide the first probabilistic characterization of AT through a unified understanding of robustness and generative ability. On the other hand, our unified framework can be extended to the unsupervised scenario, which interprets unsupervised contrastive learning as an important sampling of CEM. Based on these, we propose a principled method to develop adversarial learning and sampling methods. Experiments show that the sampling methods derived from our framework improve the sample quality in both supervised and unsupervised learning. Notably, our unsupervised adversarial sampling method achieves an Inception score of 9.61 on CIFAR-10, which is superior to previous energy-based models and comparable to state-of-the-art generative models.

****

Equivariant and Stable Positional Encoding for More Powerful Graph Neural Networks

Haorui Wang,Haoteng Yin,Muhan Zhang,Pan Li

Graph neural networks (GNN) have shown great advantages in many graph-based learning tasks but often fail to predict accurately for a task-based on sets of nodes such as link/motif prediction and so on. Many works have recently proposed to address this problem by using random node features or node distance features. However, they suffer from either slow convergence, inaccurate prediction, or high complexity. In this work, we revisit GNNs that allow using positional features of nodes given by positional encoding (PE) techniques such as Laplacian Eigenmap, Deepwalk, etc. GNNs with PE often get criticized because they are not generalizable to unseen graphs (inductive) or stable. Here, we study these issues in a principled way and propose a provable solution, a class of GNN layers termed PEG with rigorous mathematical analysis. PEG uses separate channels to update the original node features and positional features. PEG imposes permutation equivariance w.r.t. the original node features and rotation equivariance w.r.t. the positional features simultaneously. Extensive link prediction experiments over 8 real-world networks demonstrate the advantages of PEG in generalization and scalability. Code is available at https://github.com/Graph-COM/PEG.

****

Bit-wise Training of Neural Network Weights

Cristian Ivan

We propose an algorithm where the individual bits representing the weights of a neural network are learned. This method allows training weights with integer values on arbitrary bit-depths and naturally uncovers sparse networks, without additional constraints or regularization techniques. We show better results than the standard training technique with fully connected networks and similar performance as compared to standard training for residual networks. By training bits in a selective manner we found that the biggest contribution to achieving high accuracy is given by the first three most significant bits, while the rest provide an intrinsic regularization. As a consequence we show that more than 90% of a network can be used to store arbitrary codes without affecting the its accuracy. These codes can be random noise, binary files or even the weights of previously trained networks.

****

Spatial Frequency Sensitivity Regularization for Robustness

Kiran Chari,Chuan-Sheng Foo,See-Kiong Ng

The ability to generalize to out-of-distribution data is a major challenge for modern deep neural networks. Recent work has shown that deep neural networks latch on to superficial Fourier statistics of the training data and fail to generalize when these statistics change, such as when images are subject to common corruptions. In this paper, we study the frequency characteristics of deep neural networks in order to improve their robustness. We first propose a general measure of a model's $\textit{\textbf{spatial frequency sensitivity}}$ based on its input-Jacobian represented in the Fourier-basis. When applied to deep neural networks, we find that standard minibatch training consistently leads to increased sensitivity towards particular spatial frequencies independent of network architecture. We further propose a family of $\textit{\textbf{spatial frequency regularizers}}$ based on our proposed measure to induce specific spatial frequency sensitiv ities in a model. In experiments on datasets with out-of-distribution test images arising from various common image corruptions, we find that deep neural networks trained with our proposed regularizers obtain significantly improved classification accuracy while maintaining high accuracy on in-distribution clean test images.

****

Go with the Flow: the distribution of information processing in multi-path networks

Mats Leon Richter,Krupal Shah,Anna Wiedenroth,Saketh Bachu,Ulf Krumnack

The architectures of convolution neural networks (CNN) have a great impact on the predictive performance and efficiency of the model.

Yet, the development of these architectures is still driven by trial and error, making the design of novel models a costly endeavor.

To move towards a more guided process, the impact of design decisions on information processing must be understood better.

This work contributes by analyzing the processing of the information in neural architectures with parallel pathways.

Using logistic regression probes and similarity indices, we characterize the role of different pathways in the network during the inference process.

In detail, we find that similar sized pathways advance the solution quality at a similar pace, with high redundancy.

On the other hand, shorter pathways dominate longer ones by majorly transporting (and improving) the main signal, while longer pathways do not advance the solution quality directly.

Additionally, we explore the situation in which networks start to ``skip'' layers and how the skipping of layers is expressed.

****

BadPre: Task-agnostic Backdoor Attacks to Pre-trained NLP Foundation Models

Kangjie Chen,Yuxian Meng,Xiaofei Sun,Shangwei Guo,Tianwei Zhang,Jiwei Li,Chun Fan

Pre-trained Natural Language Processing (NLP) models, which can be adapted to a variety of downstream language tasks via fine-tuning, highly accelerate the learning progress of NLP models. However, NLP models have been shown to be vulnerable to backdoor attacks. Previous NLP backdoor attacks mainly focus on one specific task. This limitation makes existing solutions less applicable to different NLP models which have been widely used in various tasks.

In this work, we propose BadPre, the first backdoor attack against various downstream models built based on pre-trained NLP models. BadPre can launch trojan attacks against different language tasks with the same trigger.

The key insight of our approach is that downstream models can inherit the security characteristics from the pre-trained models. Specifically, we leverage data posing to the pre-trained NLP models and then inference the downstream models with sentences embedded triggers. Furthermore, to fool backdoor detectors, we design a novel adversarial attack method to generate a more robust trigger.

Experimental results indicate that our approach can effectively attack a wide range of downstream NLP tasks and exhibit significant robustness against backdoor detectors.

****

When in Doubt, Summon the Titans: A Framework for Efficient Inference with Large Models

Ankit Singh Rawat,Manzil Zaheer,Aditya Krishna Menon,Amr Ahmed,Sanjiv Kumar

Scaling neural networks to "large" sizes, with billions of parameters, has been shown to yield impressive results on many challenging problems. However, the inference cost incurred by such large models often prevent their application in most real-world settings. In this paper, we propose a two-stage framework based on distillation that realizes the modelling benefits of the large models, while largely preserving the computational benefits of inference with more lightweight models. In a nutshell, we use the large teacher models to guide the lightweight student models to only make correct predictions on a subset of "easy" examples; for the "hard" examples, we fall-back to the teacher. Such an approach allows us t o efficiently employ large models in practical scenarios where easy examples are much more frequent than rare hard examples. Our proposed use of distillation to only handle easy instances allows for a more aggressive trade-off in the student size, thereby reducing the amortized cost of inference and achieving better accuracy than standard distillation. Empirically, we demonstrate the benefits of our approach on both image classification and natural language processing benchmarks.

****

Shallow and Deep Networks are Near-Optimal Approximators of Korobov Functions
Moise Blanchard,Mohammed Amine Bennouna

In this paper, we analyze the number of neurons and training parameters that a neural network needs to approximate multivariate functions of bounded second mixed derivatives --- Korobov functions. We prove upper bounds on these quantities for shallow and deep neural networks, drastically lessening the curse of dimensionality. Our bounds hold for general activation functions, including ReLU. We further prove that these bounds nearly match the minimal number of parameters any continuous function approximator needs to approximate Korobov functions, showing that neural networks are near-optimal function approximators.

****

Protect the weak: Class focused online learning for adversarial training
Thomas Pethick,Grigorios Chrysos,Volkan Cevher

Adversarial training promises a defense against adversarial perturbations in terms of average accuracy. In this work, we identify that the focus on the average accuracy metric can create vulnerabilities to the "weakest" class. For instance, on CIFAR10, where the average accuracy is 47%, the worst class accuracy can be as low as 14%. The performance sacrifice of the weakest class can be detrimental for real-world systems, if indeed the threat model can adversarially choose the class to attack. To this end, we propose to explicitly minimize the worst class error, which results in a min-max-max optimization formulation. We provide high probability convergence guarantees of the worst class loss for our method, dubbed as class focused online learning (CFOL), which can be plugged into existing training setups with virtually no overhead in computation. We observe significant improvements on the worst class accuracy of 30% for CIFAR10. We also observe consistent behavior across CIFAR100 and STL10. Intriugingly, we find that minimizing the worst case can even sometimes improve the average.

****

CSQ: Centered Symmetric Quantization for Extremely Low Bit Neural Networks
Faaiz Asim,Jaewoo Park,Azat Azamat,Jongeun Lee

Recent advances in quantized neural networks (QNNs) are closing the performance gap with the full precision neural networks. However at very low precision (i.e., ≤ 3-bits), QNNs often still suffer significant performance degradation. The conventional uniform symmetric quantization scheme allocates unequal numbers of positive and negative quantization levels. We show that this asymmetry in the number of positive and negative quantization levels can result in significant quantization error and performance degradation at low precision. We propose and analyze a quantizer called centered symmetric quantizer (CSQ), which preserves the symmetry of latent distribution by providing equal representations to the negative and positive sides of the distribution. We also propose a novel method to efficiently map CSQ to binarized neural network hardware using bitwise operations. Our analyses and experimental results using state-of-the-art quantization methods on ImageNet and CIFAR-10 show the importance of using CSQ for weight in place of the conventional quantization scheme at extremely low-bit precision (2$\sim$3 bits).

****

Closed-Loop Control of Additive Manufacturing via Reinforcement Learning
Michal Piovarci,Michael Foshey,Timothy Erps,Jie Xu,Vahid Babaei,Piotr Didyk,Wojciech Matusik,Szymon Rusinkiewicz,Bernd Bickel

Additive manufacturing suffers from imperfections in hardware control and material consistency. As a result, the deposition of a large range of materials requires on-the-fly adjustment of process parameters. Unfortunately, learning the in-p rocess control is challenging. The deposition parameters are complex and highly coupled, artifacts occur after long time horizons, available simulators lack predictive power, and learning on hardware is intractable. In this work, we demonstrate the feasibility of learning a closed-loop control policy for additive manufacturing. To achieve this goal, we assume that the perception of a deposition device is limited and can capture the process only qualitatively. We leverage this assumption to formulate an efficient numerical model that explicitly includes printing imperfections. We further show that in combination with reinforcement learning, our model can be used to discover control policies that outperform state-of-the-art controllers. Furthermore, the recovered policies have a minimal sim-to-real gap. We showcase this by implementing a first-of-its-kind self-correcting printer.

****

Fast and Reliable Evaluation of Adversarial Robustness with Minimum-Margin Attack

Ruize Gao,Jiongxiao Wang,Kaiwen Zhou,Feng Liu,Binghui Xie,Gang Niu,Bo Han,James Cheng

The AutoAttack (AA) has been the most reliable method to evaluate adversarial robustness when considerable computational resources are available. However, the high computational cost (e.g., 100 times more than that of the project gradient descent attack) makes AA infeasible for practitioners with limited computational resources, and also hinders applications of AA in the adversarial training (AT). In this paper, we propose a novel method, minimum-margin (MM) attack, to fast and reliably evaluate adversarial robustness. Compared with AA, our method achieves comparable performance but only costs 3% of the computational time in extensive experiments. The reliability of our method lies in that we evaluate the quality of adversarial examples using the margin between two targets that can precisely identify the most adversarial example. The computational efficiency of our method lies in an effective Sequential TArget Ranking Selection (STARS) method, ensuring that the cost of the MM attack is independent of the number of classes. The MM attack opens a new way for evaluating adversarial robustness and contributes a feasible and reliable method to generate high-quality adversarial examples in AT.

****

What Makes Better Augmentation Strategies? Augment Difficult but Not too Different

Jaehyung Kim,Dongyeop Kang,Sungsoo Ahn,Jinwoo Shin

The practice of data augmentation has been extensively used to boost the performance of deep neural networks for various NLP tasks. It is more effective when only a limited number of labeled samples is available, e.g., low-data or class-imbalanced regimes. Most current augmentation techniques rely on parameter tuning or inherent randomness; hence, their effectiveness largely varies on the tasks. To efficiently find the best augmentation strategy for each task, learning data augmentation policy is a promising solution, but the question of what makes a good augmentation in NLP tasks and how to design the reward function for learning a good policy remains under-explored. To answer this, we hypothesize that good data augmentation should construct more diverse and challenging samples for providing informative training signals, while avoiding the risk of losing the semantics of original samples. Therefore, we design a novel reward function for updating the augmentation policy to construct difficult but not too different samples (DND). Particularly, we jointly optimize a data augmentation policy while training the model, to construct the augmented samples with low confidence but a high semantic similarity with original ones. In addition, we introduce a sample re-weighting scheme to focus on difficult augmented samples after the original ones are learned confidently for more effective learning from the augmented ones. Our learning-based augmentation outperforms the recent state-of-the-art augmentation schemes on various text classification tasks and GLUE benchmark by successfully discovering the effective augmentations for each task. Remarkably, our method is more effective on the challenging low-data and class-imbalanced regimes, and the learned augmentation policy is well-transferable to the different tasks and mod els.

****

Escaping limit cycles: Global convergence for constrained nonconvex-nonconcave minimax problems

Thomas Pethick,Puya Latafat,Panos Patrinos,Olivier Fercoq,Volkan Cevher

This paper introduces a new extragradient-type algorithm for a class of nonconvex-nonconcave minimax problems. It is well-known that finding a local solution for general minimax problems is computationally intractable. This observation has recently motivated the study of structures sufficient for convergence of first order methods in the more general setting of variational inequalities when the so-called weak Minty variational inequality (MVI) holds. This problem class captures non-trivial structures as we demonstrate with examples, for which a large family of existing algorithms provably converge to limit cycles. Our results require a less restrictive parameter range in the weak MVI compared to what is previously known, thus extending the applicability of our scheme. The proposed algorithm is applicable to constrained and regularized problems, and involves an adaptive stepsize allowing for potentially larger stepsizes. Our scheme also converges globally even in settings where the underlying operator exhibits limit cycles.

****

The Connection between Out-of-Distribution Generalization and Privacy of ML Models

Divyat Mahajan,Shruti Tople,Amit Sharma

With the goal of generalizing to out-of-distribution (OOD) data, recent domain generalization methods aim to learn ``stable'' feature representations whose effect on the output remains invariant across domains. Given the theoretical connection between generalization and privacy, we ask whether better OOD generalization leads to better privacy for machine learning models, where privacy is measured through robustness to membership inference (MI) attacks. In general, we find that the relationship does not hold. Through extensive evaluation on a synthetic dataset and image datasets like MNIST, Fashion-MNIST, and Chest X-rays, we show that a lower OOD generalization gap does not imply better robustness to MI attacks. Instead, privacy benefits are based on the extent to which a model captures the stable features. A model that captures stable features is more robust to MI attacks than models that exhibit better OOD generalization but do not learn stable features. Further, for the same provable differential privacy guarantees, a model that learns stable features provides higher utility as compared to others. Our results offer the first extensive empirical study connecting stable features and privacy, and also have a takeaway for the domain generalization community; MI attack can be used as a complementary metric to measure model quality.

****

Generative Pseudo-Inverse Memory

Kha Pham,Hung Le,Man Ngo,Truyen Tran,Bao Ho,Svetha Venkatesh

We propose Generative Pseudo-Inverse Memory (GPM), a class of deep generative memory models that are fast to write in and read out. Memory operations are recast as seeking robust solutions of linear systems, which naturally lead to the use of matrix pseudo-inverses. The pseudo-inverses are iteratively approximated, with practical computation complexity of almost $O(1)$. We prove theoretically and verify empirically that our model can retrieve exactly what have been written to the memory under mild conditions. A key capability of GPM is iterative reading, during which the attractor dynamics towards fixed points are enabled, allowing the model to iteratively improve sample quality in denoising and generating. More impressively, GPM can store a large amount of data while maintaining key abilities of accurate retrieving of stored patterns, denoising of corrupted data and generating novel samples. Empirically we demonstrate the efficiency and versatility of GPM on a comprehensive suite of experiments involving binarized MNIST, binarized Omniglot, FashionMNIST, CIFAR10 & CIFAR100 and CelebA.

****

A Deep Variational Approach to Clustering Survival Data

Laura Manduchi,Ri■ards Marcinkevi■s,Michela C. Massi,Thomas Weikert,Alexander Sauter,Verena Gotta,Timothy Müller,Flavio Vasella,Marian C. Neidert,Marc Pfister,B ram Stieltjes,Julia E Vogt

In this work, we study the problem of clustering survival data — a challenging and so far under-explored task. We introduce a novel semi-supervised probabilistic approach to cluster survival data by leveraging recent advances in stochastic gradient variational inference. In contrast to previous work, our proposed method employs a deep generative model to uncover the underlying distribution of both the explanatory variables and censored survival times. We compare our model to the related work on clustering and mixture models for survival data in comprehensive experiments on a wide range of synthetic, semi-synthetic, and real-world datasets, including medical imaging data. Our method performs better at identifying clusters and is competitive at predicting survival times. Relying on novel generative assumptions, the proposed model offers a holistic perspective on clustering survival data and holds a promise of discovering subpopulations whose survival is regulated by different generative mechanisms.

****

A New Perspective on Fluid Simulation: An Image-to-Image Translation Task via Neural Networks

Roman Lehmann,Markus Hoffmann,Simon Leufen,Wolfgang Karl

Standard numerical methods for creating simulation models in the field of fluid dynamics are designed to be close to perfection, which results in high computational effort and high computation times in many cases. Unfortunately, there is no mathematical way to decrease this correctness in cases where only approximate predictions are needed. For such cases, we developed an approach based on Neural Networks that is much less time-consuming but nearly as accurate as the numerical model for a human observer. We show that we can keep our results stable and nearly indistinguishable from their numerical counterparts over tenth to hundreds of time steps.

****

GPT-Critic: Offline Reinforcement Learning for End-to-End Task-Oriented Dialogue Systems

Youngsoo Jang,Jongmin Lee,Kee-Eung Kim

Training a task-oriented dialogue agent can be naturally formulated as offline reinforcement learning (RL) problem, where the agent aims to learn a conversational strategy to achieve user goals, only from a dialogue corpus. It is very challenging in terms of RL since the natural language action space is astronomical, while feasible (syntactically and semantically correct) actions are very sparse. Thus, standard RL methods easily fail and generate responses diverging from human language, even when fine-tuning a powerful pre-trained language model. In this paper, we introduce GPT-Critic, an offline RL method for task-oriented dialogue. GPT-Critic is built upon GPT-2, fine-tuning the language model through behavior cloning of the critic-guided self-generated sentences. GPT-Critic is essentially free from the issue of diverging from human language since it learns from the sentences sampled from the pre-trained language model. In the experiments, we demonstrate that our algorithm outperforms the state-of-the-art in the task-oriented dialogue benchmarks including MultiWOZ 2.0 and ConvLab.

****

On Pseudo-Labeling for Class-Mismatch Semi-Supervised Learning

Lu Han,Han-Jia Ye,De-Chuan Zhan

Semi-Supervised Learning (SSL) methods have shown superior performance when unlabeled data are drawn from the same distribution with labeled data. Among them, Pseudo-Labeling (PL) is a simple and widely used method that creates pseudo-labels for unlabeled data according to predictions of the training model itself. However, when there are unlabeled Out-Of-Distribution (OOD) data from other classes, these methods suffer from severe performance degradation and even get worse than merely training on labeled data. In this paper, we empirically analyze PL in class-mismatched SSL. We aim to answer the following questions: (1) How do OOD data influence PL? (2) What are the better pseudo-labels for OOD data? First, we show that the major problem of PL is imbalanced pseudo-labels on OOD data. Second, we find that when labeled as their ground truths, OOD data are beneficial to classification performance on In-Distribution (ID) data. Based on the findings, we propose our model which consists of two components -- Re-balanced Pseudo-Labeling (RPL) and Semantic Exploration Clustering (SEC). RPL re-balances pseudo-labels on ID classes to filter out OOD data while also addressing the imbalance problem. SEC uses balanced clustering on OOD data to create pseudo-labels on extra classes, simulating the process of training with their ground truths. Experiments show that our method achieves steady improvement over supervised baseline and state-of-the-art performance under all class mismatch ratios on different benchmarks.

****

Multi-Agent Language Learning: Symbolic Mapping
Yicheng Feng,Zongqing Lu

The study of emergent communication has long been devoted to coax neural network agents to learn a language sharing similar properties with human language. In this paper, we try to find a natural way to help agents learn a compositional and symmetric language in complex settings like dialog games. Inspired by the theory that human language was originated from simple interactions, we hypothesize that language may evolve from simple tasks to difficult tasks. We propose a novel architecture called symbolic mapping as a basic component of the communication system of agent. We find that symbolic mapping learned in simple referential games can notably promote language learning in difficult tasks. Further, we explore vocabulary expansion, and show that with the help of symbolic mapping, agents can easily learn to use new symbols when the environment becomes more complex. All in all, we probe into how symbolic mapping helps language learning and find that a process from simplicity to complexity can serve as a natural way to help multi-agent language learning.

****

Charformer: Fast Character Transformers via Gradient-based Subword Tokenization
Yi Tay,Vinh Q. Tran,Sebastian Ruder,Jai Gupta,Hyung Won Chung,Dara Bahri,Zhen Qin,Simon Baumgartner,Cong Yu,Donald Metzler

State-of-the-art models in natural language processing rely on separate rigid subword tokenization algorithms, which limit their generalization ability and adaptation to new settings. In this paper, we propose a new model inductive bias that learns a subword tokenization end-to-end as part of the model. To this end, we introduce a soft gradient-based subword tokenization module (GBST) that automatically learns latent subword representations from characters in a data-driven fashion. Concretely, GBST enumerates candidate subword blocks and learns to score them in a position-wise fashion using a block scoring network. We additionally introduce Charformer, a deep Transformer model that integrates GBST and operates on the character level. Via extensive experiments on English GLUE, multilingual, and noisy text datasets, we show that Charformer outperforms a series of competitive character-level baselines while generally performing on par and sometimes outperforming subword-based models. Additionally, Charformer is fast, improving the speed of vanilla character-level Transformers by up to while maintaining quality. We believe this work paves the way for highly performant token-free models that are trained completely end-to-end.

****

Self-Organized Polynomial-time Coordination Graphs
Weijun Dong,Qianlan Yang,Zhizhou Ren,Jianhao Wang,Tonghan Wang,Chongjie Zhang

Coordination graph is a promising approach to model agent collaboration in multi-agent reinforcement learning. It factorizes a large multi-agent system into a suite of overlapping groups that represent the underlying coordination dependencies. One critical challenge in this paradigm is the complexity of computing maximum-value actions for a graph-based value factorization. It refers to the decentralized constraint optimization problem (DCOP), which and whose constant-ratio approximation are NP-hard problems. To bypass this fundamental hardness, this paper proposes a novel method, named Self-Organized Polynomial-time Coordination Graphs (SOP-CG), which uses structured graph classes to guarantee the optimality of the induced DCOPs with sufficient function expressiveness. We extend the graph topology to be state-dependent, formulate the graph selection as an imaginary agent, and finally derive an end-to-end learning paradigm from the unified Bellman optimality equation. In experiments, we show that our approach learns interpretable graph topologies, induces effective coordination, and improves performance across a variety of cooperative multi-agent tasks.

****

M6-10T: A Sharing-Delinking Paradigm for Efficient Multi-Trillion Parameter Pretraining

Junyang Lin,An Yang,Jinze Bai,Chang Zhou,Le Jiang,Xianyan Jia,Ang Wang,Jie Zhang,Yong Li,Wei Lin,Jingren Zhou,Hongxia Yang

Recent expeditious developments in deep learning algorithms, distributed training, and even hardware design for large models have enabled training extreme-scale models, say GPT-3 and Switch Transformer possessing hundreds of billions or even trillions of parameters. However, under limited resources, extreme-scale model training that requires enormous amounts of computes and memory footprint suffers from frustratingly low efficiency in model convergence. In this paper, we propose a simple training strategy called “Pseudo-to-Real” for high-memory-footprint-required large models. Pseudo-to-Real is compatible with large models with architecture of sequential layers. We demonstrate a practice of pretraining unprecedented 10-trillion-parameter model, an order of magnitude larger than the state-of-the-art, on solely 512 GPUs within 10 days. Besides demonstrating the application of Pseudo-to-Real, we also provide a technique, Granular CPU offloading, to manage CPU memory for training large model and maintain high GPU utilities. Fast training of extreme-scale models on a decent amount of resources can bring much smaller carbon footprint and contribute to greener AI.

****

Meta-Forecasting by combining Global Deep Representations with Local Adaptation

Riccardo Grazzi,Valentin Flunkert,David Salinas,Tim Januschowski,Matthias Seeger,Cedric Archambeau

While classical time series forecasting considers individual time series in isolation, recent advances based on deep learning showed that jointly learning from a large pool of related time series can boost the forecasting accuracy. However, the accuracy of these methods suffers greatly when modeling out-of-sample time series, significantly limiting their applicability compared to classical forecasting methods. To bridge this gap, we adopt a meta-learning view of the time series forecasting problem. We introduce a novel forecasting method, called Meta Global- Local Auto-Regression (Meta-GLAR), that adapts to each time series by learning in closed-form the mapping from the representations produced by a recurrent neural network (RNN) to one-step-ahead forecasts. Crucially, the parameters of the RNN are learned across multiple time series by backpropagating through the closed-form adaptation mechanism. In our extensive empirical evaluation we show that our method is competitive with the state-of-the-art in out-of-sample forecasting accuracy reported in earlier work.

****

Regularized Autoencoders for Isometric Representation Learning

Yonghyeon LEE,Sangwoong Yoon,MinJun Son,Frank C. Park

The recent success of autoencoders for representation learning can be traced in large part to the addition of a regularization term.

Such regularized autoencoders ``constrain" the representation so as to prevent overfitting to the data while producing a parsimonious generative model. A regularized autoencoder should in principle learn not only the data manifold, but also a set of geometry-preserving coordinates for the latent representation space; by geometry-preserving we mean that the latent space representation should attempt to preserve actual distances and angles on the data manifold. In this paper we first formulate a hierarchy for geometry-preserving mappings (isometry, conformal mapping of degree k, area-preserving mappings). We then show that a conformal regularization term of degree zero -- i.e., one that attempts to preserve angles and relative distances, instead of angles and exact distances -- produces data representations that are superior to other existing methods. Applying our algorithm to an unsupervised information retrieval task for CelebA data with 40 annotations, we achieve 79\% precision at five retrieved images, an improvement of more than 10\% compared to recent related work. Code is available at https://git

hub.com/Gabe-YHLee/IRVAE-public.
****

Knowledge Removal in Sampling-based Bayesian Inference
Shaopeng Fu,Fengxiang He,Dacheng Tao
The right to be forgotten has been legislated in many countries, but its enforcement in the AI industry would cause unbearable costs. When single data deletion requests come, companies may need to delete the whole models learned with massive resources. Existing works propose methods to remove knowledge learned from data for explicitly parameterized models, which however are not appliable to the sampling-based Bayesian inference, {\it i.e.}, Markov chain Monte Carlo (MCMC), as MCMC can only infer implicit distributions. In this paper, we propose the first machine unlearning algorithm for MCMC. We first convert the MCMC unlearning problem into an explicit optimization problem. Based on this problem conversion, an {\it MCMC influence function} is designed to provably characterize the learned knowledge from data, which then delivers the MCMC unlearning algorithm. Theoretical analysis shows that MCMC unlearning would not compromise the generalizability of the MCMC models. Experiments on Gaussian mixture models and Bayesian neural networks confirm the effectiveness of the proposed algorithm. The code is available at \url{https://github.com/fshp971/mcmc-unlearning}.
****

Actor-critic is implicitly biased towards high entropy optimal policies
Yuzheng Hu,Ziwei Ji,Matus Telgarsky
We show that the simplest actor-critic method — a linear softmax policy updated with TD through interaction with a linear MDP, but featuring no explicit regularization or exploration — does not merely find an optimal policy, but moreover prefers high entropy optimal policies. To demonstrate the strength of this bias, the algorithm not only has no regularization, no projections, and no exploration like ϵ-greedy, but is moreover trained on a single trajectory with no resets. The key consequence of the high entropy bias is that uniform mixing assumptions on the MDP, which exist in some form in all prior work, can be dropped: the implicit regularization of the high entropy bias is enough to ensure that all chains mix and an optimal policy is reached with high probability. As auxiliary contributions, this work decouples concerns between the actor and critic by writing the actor update as an explicit mirror descent, provides tools to uniformly bound mixing times within KL balls of policy space, and provides a projection-free TD analysis with its own implicit bias which can be run from an unmixed starting distribution.

****

Igeood: An Information Geometry Approach to Out-of-Distribution Detection
Eduardo Dadalto Camara Gomes,Florence Alberge,Pierre Duhamel,Pablo Piantanida
Reliable out-of-distribution (OOD) detection is fundamental to implementing safer modern machine learning (ML) systems. In this paper, we introduce Igeood, an effective method for detecting OOD samples. Igeood applies to any pre-trained neural network, works under various degrees of access to the ML model, does not require OOD samples or assumptions on the OOD data but can also benefit (if available) from OOD samples. By building on the geodesic (Fisher-Rao) distance between the underlying data distributions, our discriminator can combine confidence scores from the logits outputs and the learned features of a deep neural network. Empirically, we show that Igeood outperforms competing state-of-the-art methods on a variety of network architectures and datasets.
****

A Fair Generative Model Using Total Variation Distance
Soobin Um,Changho Suh
We explore a fairness-related challenge that arises in generative models. The challenge is that biased training data with imbalanced representations of demographic groups may yield a high asymmetry in size of generated samples across distinct groups. We focus on practically-relevant scenarios wherein demographic labels are not available and therefore the design of a fair generative model is particularly challenging. In this paper, we propose an optimization framework that reg

ulates such unfairness by employing one prominent statistical notion, total variation distance (TVD). We quantify the degree of unfairness via the TVD between the generated samples and balanced-yet-small reference samples. We take a variational optimization approach to faithfully implement the TVD-based measure. Experiments on benchmark real datasets demonstrate that the proposed framework can significantly improve the fairness performance while maintaining realistic sample quality for a wide range of the reference set size all the way down to 1% relative to training set.

****

Bag of Instances Aggregation Boosts Self-supervised Distillation

Haohang Xu,Jiemin Fang,XIAOPENG ZHANG,Lingxi Xie,Xinggang Wang,Wenrui Dai,Hongkai Xiong,Qi Tian

Recent advances in self-supervised learning have experienced remarkable progress, especially for contrastive learning based methods, which regard each image as well as its augmentations as an individual class and try to distinguish them from all other images. However, due to the large quantity of exemplars, this kind of pretext task intrinsically suffers from slow convergence and is hard for optimization. This is especially true for small-scale models, in which we find the performance drops dramatically comparing with its supervised counterpart. In this paper, we propose a simple but effective distillation strategy for unsupervised learning. The highlight is that the relationship among similar samples counts and can be seamlessly transferred to the student to boost the performance. Our method, termed as BINGO, which is short for Bag of InstaNces aGgregatiOn, targets at transferring the relationship learned by the teacher to the student. Here bag of instances indicates a set of similar samples constructed by the teacher and are grouped within a bag, and the goal of distillation is to aggregate compact representations over the student with respect to instances in a bag. Notably, BINGO achieves new state-of-the-art performance on small-scale models, i.e., 65.5% and 68.9% top-1 accuracies with linear evaluation on ImageNet, using ResNet-18 and ResNet-34 as the backbones respectively, surpassing baselines (52.5% and 57.4% top-1 accuracies) by a significant margin. The code is available at https://github.com/haohang96/bingo.

****

ViViT: Curvature access through the generalized Gauss-Newton's low-rank structure

Felix Dangel,Lukas Tatzel,Philipp Hennig

Curvature in form of the Hessian or its generalized Gauss-Newton (GGN) approximation is valuable for algorithms that rely on a local model for the loss to train, compress, or explain deep networks. Existing methods based on implicit multiplication via automatic differentiation or Kronecker-factored block diagonal approximations do not consider noise in the mini-batch. We present ViViT, a curvature model that leverages the GGN's low-rank structure without further approximations. It allows for efficient computation of eigenvalues, eigenvectors, as well as per-sample first- and second-order directional derivatives. The representation is computed in parallel with gradients in one backward pass and offers a fine-grained cost-accuracy trade-off, which allows it to scale. As examples for ViViT's usefulness, we investigate the directional first- and second-order derivatives during training, and how noise information can be used to improve the stability of second-order methods.

****

Stability Regularization for Discrete Representation Learning

Adeel Pervez,Efstratios Gavves

We present a method for training neural network models with discrete stochastic variables.

The core of the method is \emph{stability regularization}, which is a regularization procedure based on the idea of noise stability developed in Gaussian isoperimetric theory in the analysis of Gaussian functions.

Stability regularization is method to make the output of continuous functions of Gaussian random variables close to discrete, that is binary or categorical, without the need for significant manual tuning.

The method allows control over the extent to which a Gaussian function's output is close to discrete, thus allowing for continued flow of gradient.
The method can be used standalone or in combination with existing continuous relaxation methods.
We validate the method in a broad range of experiments using discrete variables including neural relational inference, generative modeling, clustering and conditional computing.
****
Unrolling PALM for Sparse Semi-Blind Source Separation
Mohammad Fahes,Christophe Kervazo,Jérôme Bobin,Florence Tupin
Sparse Blind Source Separation (BSS) has become a well established tool for a wide range of applications – for instance, in astrophysics and remote sensing. Classical sparse BSS methods, such as the Proximal Alternating Linearized Minimization (PALM) algorithm, nevertheless often suffer from a difficult hyper-parameter choice, which undermines their results. To bypass this pitfall, we propose in this work to build on the thriving field of algorithm unfolding/unrolling. Unrolling PALM enables to leverage the data-driven knowledge stemming from realistic simulations or ground-truth data by learning both PALM hyper-parameters and variables. In contrast to most existing unrolled algorithms, which assume a fixed known dictionary during the training and testing phases, this article further emphasizes on the ability to deal with variable mixing matrices (a.k.a. dictionaries). The proposed Learned PALM (LPALM) algorithm thus enables to perform semi-blind source separation, which is key to increase the generalization of the learnt model in real-world applications. We illustrate the relevance of LPALM in astrophysical multispectral imaging: the algorithm not only needs up to $10^4–10^5$ times less iterations than PALM, but also improves the separation quality, while avoiding the cumbersome hyper-parameter and initialization choice of PALM. We further show that LPALM outperforms other unrolled source separation methods in the semi-blind setting.
****
Boosting the Transferability of Adversarial Attacks with Reverse Adversarial Perturbation
Zeyu Qin,Yanbo Fan,Yi Liu,Yong Zhang,Jue Wang,Baoyuan Wu
Deep neural networks (DNNs) have 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 of 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. Inspired by the connection between the flatness of loss landscape and the model generalization, we propose a novel attack method, dubbed reverse adversarial perturbation (RAP) to boost the transferability of adversarial examples. Specifically, instead of purely minimizing the adversarial loss at a single adversarial point, we advocate seeking adversarial examples locating at the low-value and flat region of the loss landscape, through 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. Comprehensive experimental comparisons demonstrate that RAP can significantly boost the 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.
****
Fast Generic Interaction Detection for Model Interpretability and Compression
Tianjian Zhang,Feng Yin,Zhi-Quan Luo
The ability of discovering feature interactions in a black-box model is vital to explainable deep learning. We propose a principled, global interaction detection method by casting our target as a multi-arm bandits problem and solving it swi

ftly with the UCB algorithm. This adaptive method is free of ad-hoc assumptions and among the cutting-edge methods with outstanding detection accuracy and stability. Based on the detection outcome, a lightweight and interpretable deep learning model (called ParaACE) is further built using the alternating conditional expectation (ACE) method. Our proposed ParaACE improves the prediction performance by 26 % and reduces the model size by 100+ times as compared to its Teacher model over various datasets. Furthermore, we show the great potential of our method for scientific discovery through interpreting various real datasets in the economics and smart medicine sectors. The code is available at https://github.com/zhangtj1996/ParaACE.

****

What classifiers know what they don't know?

Mohamed Ishmael Belghazi,David Lopez-Paz

Being uncertain when facing the unknown is key to intelligent decision making. However, machine learning algorithms lack reliable estimates about their predictive uncertainty. This leads to wrong and overly-confident decisions when encountering classes unseen during training. Despite the importance of equipping classifiers with uncertainty estimates ready for the real world, prior work has focused on small datasets and little or no class discrepancy between training and testing data. To close this gap, we introduce UIMNET: a realistic, ImageNet-scale test-bed to evaluate predictive uncertainty estimates for deep image classifiers. Our benchmark provides implementations of eight state-of-the-art algorithms, six uncertainty measures, four in-domain metrics, three out-domain metrics, and a fully automated pipeline to train, calibrate, ensemble, select, and evaluate models. Our test-bed is open-source and all of our results are reproducible from a fixed commit in our repository. Adding new datasets, algorithms, measures, or metrics is a matter of a few lines of code-in so hoping that UIMNET becomes a stepping stone towards realistic, rigorous, and reproducible research in uncertainty estimation. Our results show that ensembles of ERM classifiers as well as single MIMO classifiers are the two best alternatives currently available to measure uncertainty about both in-domain and out-domain classe.

****

Superior Performance with Diversified Strategic Control in FPS Games Using General Reinforcement Learning

Shuxing Li,Jiawei Xu,Chun Yuan,peng sun,Zhuobin Zheng,Zhengyou Zhang,Lei Han

This paper offers an overall solution for first-person shooter (FPS) games to achieve superior performance using general reinforcement learning (RL). We introduce an agent in ViZDoom that can surpass previous top agents ranked in the open ViZDoom AI Competitions by a large margin. The proposed framework consists of a number of generally applicable techniques, including hindsight experience replay (HER) based navigation, hindsight proximal policy optimization (HPPO), rule-guided policy search (RGPS), prioritized fictitious self-play (PFSP), and diversified strategic control (DSC). The proposed agent outperforms existing agents by taking advantage of diversified and human-like strategies, instead of larger neural networks, more accurate frag skills, or hand-craft tricks, etc. We provide comprehensive analysis and experiments to elaborate the effect of each component in affecting the agent performance, and demonstrate that the proposed and adopted techniques are important to achieve superior performance in general end-to-end FPS games. The proposed methods can contribute to other games and real-world tasks which also require spatial navigation and diversified behaviors.

****

Role Diversity Matters: A Study of Cooperative Training Strategies for Multi-Agent RL

Siyi Hu,Chuanlong Xie,Xiaodan Liang,Xiaojun Chang

Cooperative multi-agent reinforcement learning (MARL) is making rapid progress for solving tasks in a grid world and real-world scenarios, in which agents are given different attributes and goals. For example, in Starcraft II battle tasks, agents are initialized with the various move, defense, and attack abilities according to their unit types. Current researchers tend to treat different agents equally and expect them to form a joint policy automatically. However, ignoring th e differences between agents in these scenarios may bring policy degradation. Accordingly, in this study, we quantify the agent's difference and study the relationship between the agent's role and the model performance via {\bf Role Diversity}, a metric that can describe MARL tasks. We define role diversity from three perspectives: policy-based, trajectory-based, and contribution-based to fully describe the agents' differences. Through theoretical analysis, we find that the error bound in MARL can be decomposed into three parts that have a strong relation to the role diversity. The decomposed factors can significantly impact policy optimization on parameter sharing, communication mechanism, and credit assignment strategy. Role diversity can therefore serve as a flag for selecting a suitable training strategy and helping to avoid possible bottlenecks on current tasks. The main experimental platforms are based on {\bf Multiagent Particle Environment (MPE) }and {\bf The StarCraft Multi-Agent Challenge (SMAC)}, with extensions to ensure the requirement of this study are met. Our experimental results clearly show that role diversity can serve as a robust description for the characteristics of a multi-agent cooperation task and help explain the question of why the performance of different MARL training strategies is unstable according to this description. In addition, role diversity can help to find a better training strategy and increase performance in cooperative MARL.

****

Trident Pyramid Networks: The importance of processing at the feature pyramid level for better object detection

Cédric Picron,Tinne Tuytelaars

Feature pyramids have become ubiquitous in multi-scale computer vision tasks such as object detection. Based on their importance, we divide a computer vision network into three parts: a backbone (generating a feature pyramid), a core (refining the feature pyramid) and a head (generating the final output). Most existing networks operating on feature pyramids, named cores, are shallow and mostly focus on communication-based processing in the form of top-down and bottom-up operations. We present a new core architecture called Trident Pyramid Network (TPN), that allows for a deeper design and for a better balance between communication-based processing and self-processing. We show consistent improvements when using our TPN core on the COCO object detection benchmark, outperforming the popular BiFPN baseline by 1.5 AP. Additionally, we empirically show that it is more beneficial to put additional computation into the TPN core, rather than into the backbone, by outperforming a ResNet-101+FPN baseline with our ResNet-50+TPN network by 1.7 AP, while operating under similar computation budgets. This emphasizes the importance of performing computation at the feature pyramid level in modern-day object detection systems. Code will be released.

****

Reversible Instance Normalization for Accurate Time-Series Forecasting against Distribution Shift

Taesung Kim,Jinhee Kim,Yunwon Tae,Cheonbok Park,Jang-Ho Choi,Jaegul Choo

Statistical properties such as mean and variance often change over time in time series, i.e., time-series data suffer from a distribution shift problem. This change in temporal distribution is one of the main challenges that prevent accurate time-series forecasting. To address this issue, we propose a simple yet effective normalization method called reversible instance normalization (RevIN), a generally-applicable normalization-and-denormalization method with learnable affine transformation. The proposed method is symmetrically structured to remove and restore the statistical information of a time-series instance, leading to significant performance improvements in time-series forecasting, as shown in Fig. 1. We demonstrate the effectiveness of RevIN via extensive quantitative and qualitative analyses on various real-world datasets, addressing the distribution shift problem.

****

Implicit Jacobian regularization weighted with impurity of probability output

Sungyoon Lee,Jinseong Park,Jaewook Lee

Gradient descent (GD) plays a crucial role in the success of deep learning, but it is still not fully understood how GD finds minima that generalize well. In ma ny studies, GD has been understood as a gradient flow in the limit of vanishing learning rate. However, this approach has a fundamental limitation in explaining the oscillatory behavior with iterative catapult in a practical finite learning rate regime. To address this limitation, we rather start with strong empirical evidence of the plateau of the sharpness (the top eigenvalue of the Hessian) of the loss function landscape. With this observation, we investigate the Hessian through simple and much lower-dimensional matrices. In particular, to analyze the sharpness, we instead explore the eigenvalue problem for the low-dimensional matrix which is a rank-one modification of a diagonal matrix. The eigendecomposition provides a simple relation between the eigenvalues of the low-dimensional matrix and the impurity of the probability output. We exploit this connection to derive sharpness-impurity-Jacobian relation and to explain how the sharpness influences the learning dynamics and the generalization performance. In particular, we show that GD has implicit regularization effects on the Jacobian norm weighted with the impurity of the probability output.

****

Supervised Permutation Invariant Networks for solving the CVRP with bounded fleet size

Daniela Thyssens,Jonas Falkner,Lars Schmidt-Thieme

Learning to solve combinatorial optimization problems, such as the vehicle routing problem, offers great computational advantages over classical operation research solvers and heuristics. The recently developed deep reinforcement learning approaches either improve an initially given solution iteratively or sequentially construct a set of individual tours.

However, all existing learning-based approaches are not able to work for a fixed number of vehicles and thus bypass the NP-hardness of the original problem. On the other hand, this makes them less suitable for real applications, as many logistic service providers rely on solutions provided for a specific bounded fleet size and cannot accommodate short term changes to the number of vehicles.

In contrast we propose a powerful supervised deep learning framework that constructs a complete tour plan from scratch while respecting an apriori fixed number of vehicles.

In combination with an efficient post-processing scheme, our supervised approach is not only much faster and easier to train but also achieves competitive results that incorporate the practical aspect of vehicle costs.

In thorough controlled experiments we re-evaluate and compare our method to multiple state-of-the-art approaches where we demonstrate stable performance and shed some light on existent inconsistencies in the experimentation protocols of the related work.

****

Differentiable Top-k Classification Learning

Felix Petersen,Hilde Kuehne,Christian Borgelt,Oliver Deussen

The top-k classification accuracy is one of the core metrics in machine learning. Here, k is conventionally a positive integer, such as 1 or 5. In this work, we relax this assumption and propose to draw k from a probability distribution for training. Combining this with recent advances in differentiable sorting and ranking, we propose a new family of differentiable top-k cross-entropy classification losses. We find that relaxing k does not only produce better top-5 accuracies, but also makes models more robust, which leads to top-1 accuracy improvements. When fine-tuning publicly available ImageNet models, we achieve a new state-of-the-art on ImageNet for publicly available models with an 88.36% top-1 and a 98.71% top-5 accuracy.

****

Network Learning in Quadratic Games from Fictitious Plays

KEMI DING,Yijun Chen,Lei Wang,Xiaoqiang Ren,Guodong Shi

We study the ability of an adversary learning the underlying interaction network from repeated fictitious plays in linear-quadratic games. The adversary may strategically perturb the decisions for a set of action-compromised players, and observe the sequential decisions from a set of action-leaked players. Then the question lies in whether such an adversary can fully re-construct, or effectively e stimate the underlying interaction structure among the players. First of all, by drawing connections between this network learning problem in games and classical system identification theory, we establish a series of results characterizing the learnability of the interaction graph from the adversary's point of view. Next, in view of the inherent stability and sparsity constraints for the network interaction structure, we propose a stable and sparse system identification framework for learning the interaction graph from full player action observations. We also propose a stable and sparse subspace identification framework for learning the interaction graph from partially observed player actions. Finally, the effectiveness of the proposed learning frameworks is demonstrated in numerical examples.

****

Enhancing semi-supervised learning via self-interested coalitional learning

Huiling Qin,Xianyuan Zhan,Yuanxun li,Haoran Xu,yu zheng

Semi-supervised learning holds great promise for many real-world applications, due to its ability to leverage both unlabeled and expensive labeled data. However, most semi-supervised learning algorithms still heavily rely on the limited labeled data to infer and utilize the hidden information from unlabeled data. We note that any semi-supervised learning task under the self-training paradigm also hides an auxiliary task of discriminating label observability. Jointly solving these two tasks allows full utilization of information from both labeled and unlabeled data, thus alleviating the problem of over-reliance on labeled data. This naturally leads to a new learning framework, which we call Self-interested Coalitional Learning (SCL). The key idea of SCL is to construct a semi-cooperative ``game”, which forges cooperation between a main self-interested semi-supervised learning task and a companion task that infers label observability to facilitate main task training. We show with theoretical deduction its connection to loss reweighting on noisy labels. Through comprehensive evaluation on both classification and regression tasks, we show that SCL can consistently enhance the performance of semi-supervised learning algorithms.

****

The Effect of diversity in Meta-Learning

Ramnath Kumar,Tristan Deleu,Yoshua Bengio

Few-shot learning aims to learn representations that can tackle novel tasks given a small number of examples. Recent studies show that task distribution plays a vital role in the performance of the model. Conventional wisdom is that task diversity should improve the performance of meta-learning. In this work, we find evidence to the contrary; we study different task distributions on a myriad of models and datasets to evaluate the effect of task diversity on meta-learning algorithms. For this experiment, we train on multiple datasets, and with three broad classes of meta-learning models - Metric-based (i.e., Protonet, Matching Networks), Optimization-based (i.e., MAML, Reptile, and MetaOptNet), and Bayesian meta-learning models (i.e., CNAPs). Our experiments demonstrate that the effect of task diversity on all these algorithms follows a similar trend, and task diversity does not seem to offer any benefits to the learning of the model. Furthermore, we also demonstrate that even a handful of tasks, repeated over multiple batches, would be sufficient to achieve a performance similar to uniform sampling and draws into question the need for additional tasks to create better models.

****

On the Pitfalls of Analyzing Individual Neurons in Language Models

Omer Antverg,Yonatan Belinkov

While many studies have shown that linguistic information is encoded in hidden word representations, few have studied individual neurons, to show how and in which neurons it is encoded.

Among these, the common approach is to use an external probe to rank neurons according to their relevance to some linguistic attribute, and to evaluate the obtained ranking using the same probe that produced it.

We show two pitfalls in this methodology:

1. It confounds distinct factors: probe quality and ranking quality.

We separate them and draw conclusions on each.

2. It focuses on encoded information, rather than information that is used by the model.

We show that these are not the same.

We compare two recent ranking methods and a simple one we introduce, and evaluate them with regard to both of these aspects.

****

Query Embedding on Hyper-Relational Knowledge Graphs

Dimitrios Alivanistos,Max Berrendorf,Michael Cochez,Mikhail Galkin

Multi-hop logical reasoning is an established problem in the field of representation learning on knowledge graphs (KGs). It subsumes both one-hop link prediction as well as other more complex types of logical queries. Existing algorithms operate only on classical, triple-based graphs, whereas modern KGs often employ a hyper-relational modeling paradigm. In this paradigm, typed edges may have several key-value pairs known as qualifiers that provide fine-grained context for facts. In queries, this context modifies the meaning of relations, and usually reduces the answer set. Hyper-relational queries are often observed in real-world KG applications, and existing approaches for approximate query answering cannot make use of qualifier pairs. In this work, we bridge this gap and extend the multi-hop reasoning problem to hyper-relational KGs allowing to tackle this new type of complex queries. Building upon recent advancements in Graph Neural Networks and query embedding techniques, we study how to embed and answer hyper-relational conjunctive queries. Besides that, we propose a method to answer such queries and demonstrate in our experiments that qualifiers improve query answering on a diverse set of query patterns.

****

Neural Solvers for Fast and Accurate Numerical Optimal Control

Federico Berto,Stefano Massaroli,Michael Poli,Jinkyoo Park

Synthesizing optimal controllers for dynamical systems often involves solving optimization problems with hard real-time constraints. These constraints determine the class of numerical methods that can be applied: computationally expensive but accurate numerical routines are replaced by fast and inaccurate methods, trading inference time for solution accuracy. This paper provides techniques to improve the quality of optimized control policies given a fixed computational budget. We achieve the above via a hypersolvers approach, which hybridizes a differential equation solver and a neural network. The performance is evaluated in direct and receding-horizon optimal control tasks in both low and high dimensions, where the proposed approach shows consistent Pareto improvements in solution accuracy and control performance.

****

Graph Barlow Twins: A self-supervised representation learning framework for graphs

Piotr Bielak,Tomasz Jan Kajdanowicz,Nitesh Chawla

The self-supervised learning (SSL) paradigm is an essential exploration area, which tries to eliminate the need for expensive data labeling. Despite the great success of SSL methods in computer vision and natural language processing, most of them employ contrastive learning objectives that require negative samples, which are hard to define. This becomes even more challenging in the case of graphs and is a bottleneck for achieving robust representations. To overcome such limitations, we propose a framework for self-supervised graph representation learning - Graph Barlow Twins, which utilizes a cross-correlation-based loss function instead of negative samples. Moreover, it does not rely on non-symmetric neural network architectures - in contrast to state-of-the-art self-supervised graph representation learning method BGRL. We show that our method achieves as competitive results as the best self-supervised methods and fully supervised ones while requiring fewer hyperparameters and substantially shorter computation time (ca. 30 times faster than BGRL).

****

ON THE GENERALIZATION OF WASSERSTEIN ROBUST FEDERATED LEARNING

Long Tan Le,Josh Nguyen,Canh T. Dinh,Nguyen Hoang Tran

In Federated learning (FL), participating clients typically possess non-i.i.d. data, posing a significant challenge to generalization to unseen distributions. To address this, we propose a Wasserstein distributionally robust optimization scheme called WAFL. Leveraging its duality, we frame WAFL as an empirical surrogate risk minimization problem, and solve it using a novel local SGD-based algorithm with convergence guarantees. We show that the robustness of WAFL is more general than related approaches, and the generalization bound is robust to all adversarial distributions inside the Wasserstein ball (ambiguity set). Since the center location and radius of the Wasserstein ball can be suitably modified, WAFL shows its applicability not only in robustness but also in domain adaptation. Through empirical evaluation, we demonstrate that WAFL generalizes better than the vanilla FedAvg in non-i.i.d. settings, and is more robust than other related methods in distribution shift settings. Further, using benchmark datasets we show that WAFL is capable of generalizing to unseen target domains.
****

Disentangling deep neural networks with rectified linear units using duality
CHANDRA SHEKAR LAKSHMINARAYANAN,Amit Vikram Singh

Despite their success deep neural networks (DNNs) are still largely considered as black boxes. The main issue is that the linear and non-linear operations are entangled in every layer, making it hard to interpret the hidden layer outputs. In this paper, we look at DNNs with rectified linear units (ReLUs), and focus on the gating property (‘on/off’ states) of the ReLUs. We extend the recently developed dual view in which the computation is broken path-wise to show that learning in the gates is more crucial, and learning the weights given the gates is characterised analytically via the so called neural path kernel (NPK) which depends on inputs and gates. In this paper, we present novel results to show that convolution with global pooling and skip connection provide respectively rotational invariance and ensemble structure to the NPK. To address ‘black box’-ness, we propose a novel interpretable counterpart of DNNs with ReLUs namely deep linearly gated networks (DLGN): the pre- activations to the gates are generated by a deep linear network, and the gates are then applied as external masks to learn the weights in a different network. The DLGN is not an alternative architecture per se, but a disentanglement and an interpretable re-arrangement of the computations in a DNN with ReLUs. The DLGN disentangles the computations into two ‘mathematically’ interpretable linearities (i) the ‘primal’ linearity between the input and the pre-activations in the gating network and (ii) the ‘dual’ linearity in the path space in the weights network characterised by the NPK. We compare the performance of DNN, DGN and DLGN on CIFAR-10 and CIFAR-100 to show that, the DLGN recovers more than 83.5% of the performance of state-of-the-art DNNs, i.e., while entanglement in the DNNs enable their improved performance, the ‘disentangled and interpretable’ computations in the DLGN recovers most part of the performance. This brings us to an interesting question: ‘Is DLGN a universal spectral approximator?’
****

Compositional Attention: Disentangling Search and Retrieval
Sarthak Mittal,Sharath Chandra Raparthy,Irina Rish,Yoshua Bengio,Guillaume Lajoie

Multi-head, key-value attention is the backbone of transformer-like model architectures which have proven to be widely successful in recent years. This attention mechanism uses multiple parallel key-value attention blocks (called heads), each performing two fundamental computations: (1) search - selection of a relevant entity from a set via query-key interaction, and (2) retrieval - extraction of relevant features from the selected entity via a value matrix. Standard attention heads learn a rigid mapping between search and retrieval. In this work, we first highlight how this static nature of the pairing can potentially: (a) lead to learning of redundant parameters in certain tasks, and (b) hinder generalization. To alleviate this problem, we propose a novel attention mechanism, called Compositional Attention, that replaces the standard head structure. The proposed mechanism disentangles search and retrieval and composes them in a dynamic, flexible and context-dependent manner. Through a series of numerical experiments, we show that it outperforms standard multi-head attention on a variety of tasks, including some out-of-distribution settings. Through our qualitative analysis, we demonstrate that Compositional Attention leads to dynamic specialization based on the type of retrieval needed. Our proposed mechanism generalizes multi-head attention, allows independent scaling of search and retrieval and is easy to implement in a variety of established network architectures.

****

SLASH: Embracing Probabilistic Circuits into Neural Answer Set Programming

Arseny Skryagin,Wolfgang Stammer,Daniel Ochs,Devendra Singh Dhami,Kristian Kersting

The goal of combining the robustness of neural networks and the expressivity of symbolic methods has rekindled the interest in Neuro-Symbolic AI. Recent advancements in Neuro-Symbolic AI often consider specifically-tailored architectures consisting of disjoint neural and symbolic components, and thus do not exhibit desired gains that can be achieved by integrating them into a unifying framework. We introduce SLASH -- a novel deep probabilistic programming language (DPPL). At its core, SLASH consists of Neural-Probabilistic Predicates (NPPs) and logical programs which are united via answer set programming. The probability estimates resulting from NPPs act as the binding element between the logical program and raw input data, thereby allowing SLASH to answer task-dependent logical queries. This allows SLASH to elegantly integrate the symbolic and neural components in a unified framework. We evaluate SLASH on the benchmark data of MNIST addition as well as novel tasks for DPPLs such as missing data prediction and set prediction with state-of-the-art performance, thereby showing the effectiveness and generality of our method.

****

PSA-GAN: Progressive Self Attention GANs for Synthetic Time Series

Paul Jeha,Michael Bohlke-Schneider,Pedro Mercado,Shubham Kapoor,Rajbir Singh Nirwan,Valentin Flunkert,Jan Gasthaus,Tim Januschowski

Realistic synthetic time series data of sufficient length enables practical applications in time series modeling tasks, such as forecasting, but remains a challenge. In this paper we present PSA-GAN, a generative adversarial network (GAN) that generates long time series samples of high quality using progressive growing of GANs and self-attention. We show that PSA-GAN can be used to reduce the error in several downstream forecasting tasks over baselines that only use real data. We also introduce a Frechet-Inception Distance-like score for time series, Context-FID, assessing the quality of synthetic time series samples. We find that Context-FID is indicative for downstream performance. Therefore, Context-FID could be a useful tool to develop time series GAN models.

****

ToM2C: Target-oriented Multi-agent Communication and Cooperation with Theory of Mind

Yuanfei Wang,fangwei zhong,Jing Xu,Yizhou Wang

Being able to predict the mental states of others is a key factor to effective social interaction. It is also crucial for distributed multi-agent systems, where agents are required to communicate and cooperate. In this paper, we introduce such an important social-cognitive skill, i.e. Theory of Mind (ToM), to build socially intelligent agents who are able to communicate and cooperate effectively to accomplish challenging tasks. With ToM, each agent is capable of inferring the mental states and intentions of others according to its (local) observation. Based on the inferred states, the agents decide "when'' and with "whom'' to share their intentions. With the information observed, inferred, and received, the agents decide their sub-goals and reach a consensus among the team. In the end, the low-level executors independently take primitive actions to accomplish the sub-goals. We demonstrate the idea in two typical target-oriented multi-agent tasks: cooperative navigation and multi-sensor target coverage. The experiments show that the proposed model not only outperforms the state-of-the-art methods on reward and communication efficiency, but also shows good generalization across different scales of the environment.

****

PROMISSING: Pruning Missing Values in Neural Networks

Seyed Mostafa Kia,Nastaran Mohammadian Rad,Daniel van Opstal,Bart van Schie,Wiepke Cahn,Andre Marquand,Josien P.W. Pluim,Hugo G. Schnack

While data are the primary fuel for machine learning models, they often suffer from missing values, especially when collected in real-world scenarios. However, many off-the-shelf machine learning models, including artificial neural network models, are unable to handle these missing values directly. Therefore, extra data preprocessing and curation steps, such as data imputation, are inevitable before learning and prediction processes. In this study, we propose a simple and intuitive yet effective method for pruning missing values (PROMISSING) during learning and inference steps in neural networks. In this method, there is no need to remove or impute the missing values; instead, the missing values are treated as a new source of information (representing what we do not know). Our experiments on simulated data, several classification and regression benchmarks, and a multi-modal clinical dataset show that PROMISSING results in similar classification performance compared to various imputation techniques. In addition, our experiments show models trained using PROMISSING techniques are becoming less decisive in their predictions when facing incomplete samples with many unknowns. This finding hopefully advances machine learning models from being pure predicting machines to more realistic thinkers that can also say "I do not know" when facing incomplete sources of information.

****

Contrastive Fine-grained Class Clustering via Generative Adversarial Networks

Yunji Kim,Jung-Woo Ha

Unsupervised fine-grained class clustering is a practical yet challenging task due to the difficulty of feature representations learning of subtle object details. We introduce C3-GAN, a method that leverages the categorical inference power of InfoGAN with contrastive learning. We aim to learn feature representations that encourage a dataset to form distinct cluster boundaries in the embedding space, while also maximizing the mutual information between the latent code and its image observation. Our approach is to train a discriminator, which is also used for inferring clusters, to optimize the contrastive loss, where image-latent pairs that maximize the mutual information are considered as positive pairs and the rest as negative pairs. Specifically, we map the input of a generator, which was sampled from the categorical distribution, to the embedding space of the discriminator and let them act as a cluster centroid. In this way, C3-GAN succeeded in learning a clustering-friendly embedding space where each cluster is distinctively separable. Experimental results show that C3-GAN achieved the state-of-the-art clustering performance on four fine-grained image datasets, while also alleviating the mode collapse phenomenon. Code is available at https://github.com/naver-ai/c3-gan.

****

Approximating Instance-Dependent Noise via Instance-Confidence Embedding

Yivan Zhang,Masashi Sugiyama

Label noise in multiclass classification is a major obstacle to the deployment of learning systems. However, unlike the widely used class-conditional noise (CCN) assumption that the noisy label is independent of the input feature given the true label, label noise in real-world datasets can be aleatory and heavily dependent on individual instances. In this work, we investigate the instance-dependent noise (IDN) model and propose an efficient approximation of IDN to capture the instance-specific label corruption. Concretely, noting the fact that most columns of the IDN transition matrix have only limited influence on the class-posterior estimation, we propose a variational approximation that uses a single-scalar confidence parameter. To cope with the situation where the mapping from the instance to its confidence value could vary significantly for two adjacent instances, we suggest using instance embedding that assigns a trainable parameter to each instance. The resulting instance-confidence embedding (ICE) method not only performs well under label noise but also can effectively detect ambiguous or mislabeled instances. We validate its utility on various image and text classification tasks.

****

Robustmix: Improving Robustness by Regularizing the Frequency Bias of Deep Nets

Jonas Ngnawé,MARIANNE NJIFON,Jonathan Heek,Yann Dauphin

Deep networks have achieved impressive results on a range of well curated benchmark datasets. Surprisingly, their performance remains sensitive to perturbations that have little effect on human performance. In this work, we propose a novel extension of Mixup called Robustmix that regularizes networks to classify based on lower frequency spatial features. We show that this type of regularization improves robustness on a range of benchmarks such as Imagenet-C and Stylized Imagenet. It adds little computational overhead and furthermore does not require a priori knowledge of a large set of image transformations. We find that this approach further complements recent advances in model architecture and data augmentation attaining a state-of-the-art mCE of 44.8 with an EfficientNet-B8 model and RandAugment, which is a reduction of 16 mCE compared to the baseline.

****

Better Supervisory Signals by Observing Learning Paths

Yi Ren,Shangmin Guo,Danica J. Sutherland

Better-supervised models might have better performance. In this paper, we first clarify what makes for good supervision for a classification problem, and then explain two existing label refining methods, label smoothing and knowledge distillation, in terms of our proposed criterion. To further answer why and how better supervision emerges, we observe the learning path, i.e., the trajectory of the model's predictions during training, for each training sample. We find that the model can spontaneously refine "bad" labels through a "zig-zag" learning path, which occurs on both toy and real datasets. Observing the learning path not only provides a new perspective for understanding knowledge distillation, overfitting, and learning dynamics, but also reveals that the supervisory signal of a teacher network can be very unstable near the best points in training on real tasks. Inspired by this, we propose a new knowledge distillation scheme, Filter-KD, which improves downstream classification performance in various settings.

****

Decoupled Kernel Neural Processes: Neural Network-Parameterized Stochastic Processes using Explicit Data-driven Kernel

Daehoon Gwak,Gyubok Lee,Jaehoon Lee,Jaesik Choi,Jaegul Choo,Edward Choi

Neural Processes (NPs) are a class of stochastic processes parametrized by neural networks. Unlike traditional stochastic processes (e.g., Gaussian processes), which require specifying explicit kernel functions, NPs implicitly learn kernel functions appropriate for a given task through observed data. While this data-driven learning of stochastic processes has been shown to model various types of data, the current NPs' implicit treatment of the mean and the covariance of the output variables limits its full potential when the underlying distribution of the given data is highly complex. To address this, we introduce a new neural stochastic processes, Decoupled Kernel Neural Processes (DKNPs), which explicitly learn a separate mean and kernel function to directly model the covariance between output variables in a data-driven manner. By estimating kernel functions with self- and mixed attentive neural networks, DKNPs demonstrate improved uncertainty estimation in terms of conditional likelihood and diversity in generated samples in 1-D and 2-D regression tasks, compared to other concurrent NP variants. Also, maintaining explicit kernel functions, a key component of stochastic processes, allows the model to reveal a deeper understanding of underlying distributions.

****

DPP-TTS: Diversifying prosodic features of speech via determinantal point processes

Seongho Joo,Kyomin Jung

With the rapid advancement in deep generative models, recent neural text-to-speech models have succeeded in synthesizing human-like speech, even in an end-to-end manner. However, many synthesized samples often have a monotonous speaking style or simply follow the speaking style of their ground-truth samples. Although there have been many proposed methods to increase the diversity of prosody in spe ech, increasing prosody variance in speech often hurts the naturalness of speech. Determinantal point processes (DPPs) have shown remarkable results for modeling diversity in a wide range of machine learning tasks. However, their application in speech synthesis has not been explored. To enhance the expressiveness of speech, we propose DPP-TTS: a text-to-speech model based on a determinantal point process. The extent of prosody diversity can be easily controlled by adjusting parameters in our model. We demonstrate that DPP-TTS generates more expressive samples than baselines in the side-by-side comparison test while not harming the naturalness of the speech.

****

Explanatory Learning: Beyond Empiricism in Neural Networks

Antonio Norelli,Giorgio Mariani,Luca Moschella,Andrea Santilli,Giambattista Parascandolo,Simone Melzi,Emanuele Rodolà

We introduce Explanatory Learning (EL), an explanation-driven machine learning framework to use existing knowledge buried in symbolic sequences expressed in an unknown language. In EL, the burden of interpreting explanations is not left to humans or human-coded compilers, as done in Program Synthesis. Rather, EL calls for a learned interpreter, built upon existing explanations paired with observations of several phenomena. This interpreter can then be used to make predictions on novel phenomena, and even find an explanation for them. We formulate the EL problem as a simple binary classification task, so that common end-to-end approaches aligned with the dominant empiricist view of machine learning could, in principle, solve it. To these models, we oppose Critical Rationalist Networks (CRNs), which instead embrace a rationalist view on the acquisition of knowledge. CRNs express several desired properties by construction, they are truly explainable, can adjust their processing at test-time for harder inferences, and can offer strong confidence guarantees on their predictions.

****

Model Fusion of Heterogeneous Neural Networks via Cross-Layer Alignment

Dang Nguyen,Khai Nguyen,Nhat Ho,Dinh Phung,Hung Bui

Layer-wise model fusion via optimal transport, named OTFusion, applies soft neuron association for unifying different pre-trained networks to save computational resources. While enjoying its success, OTFusion requires the input networks to have the same number of layers. To address this issue, we propose a novel model fusion framework, named CLAFusion, to fuse neural networks with a different number of layers, which we refer to as heterogeneous neural networks, via cross-layer alignment. The cross-layer alignment problem, which is an unbalanced assignment problem, can be solved efficiently using dynamic programming. Based on the cross-layer alignment, our framework balances the number of layers of neural networks before applying layer-wise model fusion. Our synthetic experiments indicate that the fused network from CLAFusion achieves a more favorable performance compared to the individual networks trained on heterogeneous data without the need for any retraining. With an extra finetuning process, it improves the accuracy of residual networks on the CIFAR10 dataset. Finally, we explore its application for model compression and knowledge distillation when applying to the teacher-student setting.

****

Red Alarm for Pre-trained Models: Universal Vulnerability to Neuron-Level Backdoor Attacks

Zhengyan Zhang,Guangxuan Xiao,Yongwei Li,Tian Lv,Fanchao Qi,Zhiyuan Liu,Yasheng Wang,Xin Jiang,Maosong Sun

The pre-training-then-fine-tuning paradigm has been widely used in deep learning. Due to the huge computation cost for pre-training, practitioners usually download pre-trained models from the Internet and fine-tune them on downstream datasets while the downloaded models may suffer backdoor attacks. Different from previous attacks aiming at a target task, we show that a backdoored pre-trained model can behave maliciously in various downstream tasks without foreknowing task information. Attackers can restrict the output representations of trigger-embedded samples to arbitrary predefined values through additional training, namely Neuron-level Backdoor Attack (NeuBA). Since fine-tuning has little effect on model pa rameters, the fine-tuned model will retain the backdoor functionality and predict a specific label for the samples embedded with the same trigger. To provoke multiple labels in a specific task, attackers can introduce several triggers with contrastive predefined values. In the experiments of both natural language processing (NLP) and computer vision (CV), we show that NeuBA can well control the predictions for trigger-embedded instances with different trigger designs. Our findings sound a red alarm for the wide use of pre-trained models. Finally, we apply several defense methods to NeuBA and find that model pruning is a promising technique to resist NeuBA by omitting backdoored neurons.

****

Structured Uncertainty in the Observation Space of Variational Autoencoders
James Langley,Miguel Monteiro,Charles Jones,Nick Pawlowski,Ben Glocker

Variational autoencoders (VAEs) are a popular class of deep generative models with many variants and a wide range of applications. Improvements upon the standard VAE mostly focus on the modelling of the posterior distribution over the latent space and the properties of the neural network decoder. In contrast, improving the model for the observational distribution is rarely considered and typically defaults to a pixel-wise independent categorical or normal distribution. In image synthesis, sampling from such distributions produces spatially-incoherent results with uncorrelated pixel noise, resulting in only the sample mean being somewhat useful as an output prediction. In this paper, we aim to stay true to VAE theory by improving the samples from the observational distribution. We propose an alternative model for the observation space, encoding spatial dependencies via a low-rank parameterization. We demonstrate that this new observational distribution has the ability to capture relevant covariance between pixels, resulting in spatially-coherent samples. In contrast to pixel-wise independent distributions, our samples seem to contain semantically meaningful variations from the mean allowing the prediction of multiple plausible outputs with a single forward pass.

****

TAda! Temporally-Adaptive Convolutions for Video Understanding
Ziyuan Huang,Shiwei Zhang,Liang Pan,Zhiwu Qing,Mingqian Tang,Ziwei Liu,Marcelo H Ang Jr

Spatial convolutions are widely used in numerous deep video models. It fundamentally assumes spatio-temporal invariance, i.e., using shared weights for every location in different frames. This work presents Temporally-Adaptive Convolutions (TAdaConv) for video understanding, which shows that adaptive weight calibration along the temporal dimension is an efficient way to facilitate modelling complex temporal dynamics in videos. Specifically, TAdaConv empowers the spatial convolutions with temporal modelling abilities by calibrating the convolution weights for each frame according to its local and global temporal context. Compared to previous temporal modelling operations, TAdaConv is more efficient as it operates over the convolution kernels instead of the features, whose dimension is an order of magnitude smaller than the spatial resolutions. Further, the kernel calibration brings an increased model capacity. We construct TAda2D and TAdaConvNeXt networks by replacing the 2D convolutions in ResNet and ConvNeXt with TAdaConv, which leads to at least on par or better performance compared to state-of-the-art approaches on multiple video action recognition and localization benchmarks. We also demonstrate that as a readily plug-in operation with negligible computation overhead, TAdaConv can effectively improve many existing video models with a convincing margin.

****

BEiT: BERT Pre-Training of Image Transformers
Hangbo Bao,Li Dong,Songhao Piao,Furu Wei

We introduce a self-supervised vision representation model BEiT, which stands for Bidirectional Encoder representation from Image Transformers. Following BERT developed in the natural language processing area, we propose a masked image modeling task to pretrain vision Transformers. Specifically, each image has two views in our pre-training, i.e., image patches (such as 16 x 16 pixels), and visual tokens (i.e., discrete tokens). We first ``tokenize'' the original image into vi

sual tokens. Then we randomly mask some image patches and fed them into the backbone Transformer. The pre-training objective is to recover the original visual tokens based on the corrupted image patches. After pre-training BEiT, we directly fine-tune the model parameters on downstream tasks by appending task layers upon the pretrained encoder. Experimental results on image classification and semantic segmentation show that our model achieves competitive results with previous pre-training methods.
****

Model-Based Opponent Modeling
XiaoPeng Yu,Jiechuan Jiang,Haobin Jiang,Zongqing Lu

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 opponent. In this paper, we propose model-based opponent modeling (MBOM), which employs the environment model to adapt to all kinds of opponent. 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 competitive and cooperative environments, respectively with different types of opponent, i.e., fixed policy, naive learner, and reasoning learner.
****

Learning to Persuade
Xiaodong Liu,Zhikang Fan,Xun Wang,Weiran Shen

In the standard Bayesian persuasion model, an informed sender looks to design a signaling scheme to partially reveal the information to an uninformed receiver, so as to influence the behavior of the receiver. This kind of strategic interaction abounds in the real world. However, the standard model relies crucially on some stringent assumptions that usually do not hold in reality. For example, the sender knows the receiver's utility function and the receiver's behavior is completely rational.

In this paper, we aim to relax these assumptions using techniques from the AI domain. We put forward a framework that contains both a receiver model and a sender model. We first train a receiver model through interactions between the sender and the receiver. The model is used to predict the receiver's behavior when the sender's scheme changes. Then we update the sender model to obtain an approximately optimal scheme using the receiver model. Experiments show that our framework has comparable performance to the optimal scheme.
****

Learning a subspace of policies for online adaptation in Reinforcement Learning
Jean-Baptiste Gaya,Laure Soulier,Ludovic Denoyer

Deep Reinforcement Learning (RL) is mainly studied in a setting where the training and the testing environments are similar. But in many practical applications, these environments may differ. For instance, in control systems, the robot(s) on which a policy is learned might differ from the robot(s) on which a policy will run. It can be caused by different internal factors (e.g., calibration issues, system attrition, defective modules) or also by external changes (e.g., weather conditions). There is a need to develop RL methods that generalize well to variations of the training conditions. In this article, we consider the simplest yet hard to tackle generalization setting where the test environment is unknown at train time, forcing the agent to adapt to the system's new dynamics. This online adaptation process can be computationally expensive (e.g., fine-tuning) and cannot rely on meta-RL techniques since there is just a single train environment. To do so, we propose an approach where we learn a subspace of policies within the parameter space. This subspace contains an infinite number of policies that ar

e trained to solve the training environment while having different parameter values. As a consequence, two policies in that subspace process information differently and exhibit different behaviors when facing variations of the train environment. Our experiments carried out over a large variety of benchmarks compare our approach with baselines, including diversity-based methods. In comparison, our approach is simple to tune, does not need any extra component (e.g., discriminator) and learns policies able to gather a high reward on unseen environments.
****

Fine-Tuning can Distort Pretrained Features and Underperform Out-of-Distribution
Ananya Kumar,Aditi Raghunathan,Robbie Matthew Jones,Tengyu Ma,Percy Liang
When transferring a pretrained model to a downstream task, two popular methods are full fine-tuning (updating all the model parameters) and linear probing (updating only the last linear layer---the "head"). It is well known that fine-tuning leads to better accuracy in-distribution (ID). However, in this paper, we find that fine-tuning can achieve worse accuracy than linear probing out-of-distribution (OOD) when the pretrained features are good and the distribution shift is large. On 10 distribution shift datasets (BREEDS-Living17, BREEDS-Entity30, DomainNet, CIFAR $\to$ STL, CIFAR-10.1, FMoW, ImageNetV2, ImageNet-R, ImageNet-A, ImageNet-Sketch), fine-tuning obtains on average 2% higher accuracy ID but 7% lower accuracy OOD than linear probing. We show theoretically that this tradeoff between ID and OOD accuracy arises even in a simple setting: fine-tuning overparameterized two-layer linear networks. We prove that the OOD error of fine-tuning is high when we initialize with a fixed or random head---this is because while fine-tuning learns the head, the lower layers of the neural network change simultaneously and distort the pretrained features. Our analysis suggests that the easy two-step strategy of linear probing then full fine-tuning (LP-FT), sometimes used as a fine-tuning heuristic, combines the benefits of both fine-tuning and linear probing. Empirically, LP-FT outperforms both fine-tuning and linear probing on the above datasets (1% better ID, 10% better OOD than full fine-tuning).
****

TotalRecall: A Bidirectional Candidates Generation Framework for Large Scale Recommender \& Advertising Systems
Qifang Zhao,Yu Jiang,Yuqing Liu,Meng Du,Qinghui Sun,Chao Xu,huan xu,Zhongyao Wang
Recommender (RS) and Advertising/Marketing Systems (AS) play the key roles in E-commerce companies like Amazaon and Alibaba. RS needs to generate thousands of item candidates for each user ($u2i$), while AS needs to identify thousands or even millions of high-potential users for given items so that the merchant can advertise these items efficiently with limited budget ($i2u$). This paper proposes an elegant bidirectional candidates generation framework that can serve both purposes all together. Besides, our framework is also superior in these aspects: $i).$ Our framework can easily incorporate many DNN-architectures of RS ($u2i$), and increase the HitRate and Recall by a large margin. $ii).$ We archive much better results in $i2u$ candidates generation compare to strong baselines. $iii).$ We empirically show that our framework can diversify the generated candidates, and ensure fast convergence to better results.
****

ZeroFL: Efficient On-Device Training for Federated Learning with Local Sparsity
Xinchi Qiu,Javier Fernandez-Marques,Pedro PB Gusmao,Yan Gao,Titouan Parcollet,Nicholas Donald Lane
When the available hardware cannot meet the memory and compute requirements to efficiently train high performing machine learning models, a compromise in either the training quality or the model complexity is needed. In Federated Learning (FL), nodes are orders of magnitude more constrained than traditional server-grade hardware and are often battery powered, severely limiting the sophistication of models that can be trained under this paradigm. While most research has focused on designing better aggregation strategies to improve convergence rates and in alleviating the communication costs of FL, fewer efforts have been devoted to accelerating on-device training. Such stage, which repeats hundreds of times (i.e. every round) and can involve thousands of devices, accounts for the majority o

f the time required to train federated models and, the totality of the energy consumption at the client side. In this work, we present the first study on the unique aspects that arise when introducing sparsity at training time in FL workloads. We then propose ZeroFL, a framework that relies on highly sparse operations to accelerate on-device training. Models trained with ZeroFL and 95% sparsity achieve up to 2.3% higher accuracy compared to competitive baselines obtained from adapting a state-of-the-art sparse training framework to the FL setting.
****

Equivalence of State Equations from Different Methods in High-dimensional Regression
Saidi Luo,Song tao Tian,Qian Lin
State equations were firstly introduced in the approximate message passing (AMP) to describe the mean square error (MSE) in compressed sensing. Since then a set of state equations have appeared in studies of logistic regression, robust estimator and other high-dimensional statistics problems. Recently, a convex Gaussian min-max theorem(CGMT) approach was proposed to study high-dimensional statistic problems accompanying with another set of different state equations. This Paper provides a uniform viewpoint on these methods and shows the equivalence of their reduction forms, which causes that the resulting SE are essentially equivalent and can be converted into the same expression through parameter transformations. Combining these results, we show that these different state equations are derived from several equivalent reduction forms. We believe this equivalence shed light on discovering a deeper structure in high dimensional statistics.
****

Gaussian Mixture Convolution Networks
Adam Celarek,Pedro Hermosilla,Bernhard Kerbl,Timo Ropinski,Michael Wimmer
This paper proposes a novel method for deep learning based on the analytical convolution of multidimensional Gaussian mixtures.
In contrast to tensors, these do not suffer from the curse of dimensionality and allow for a compact representation, as data is only stored where details exist.
Convolution kernels and data are Gaussian mixtures with unconstrained weights, positions, and covariance matrices.
Similar to discrete convolutional networks, each convolution step produces several feature channels, represented by independent Gaussian mixtures.
Since traditional transfer functions like ReLUs do not produce Gaussian mixtures, we propose using a fitting of these functions instead.
This fitting step also acts as a pooling layer if the number of Gaussian components is reduced appropriately.
We demonstrate that networks based on this architecture reach competitive accuracy on Gaussian mixtures fitted to the MNIST and ModelNet data sets.
****

How Does SimSiam Avoid Collapse Without Negative Samples? A Unified Understanding with Self-supervised Contrastive Learning
Chaoning Zhang,Kang Zhang,Chenshuang Zhang,Trung X. Pham,Chang D. Yoo,In So Kweon
To avoid collapse in self-supervised learning (SSL), a contrastive loss is widely used but often requires a large number of negative samples. Without negative samples yet achieving competitive performance, a recent work~\citep{chen2021exploring} has attracted significant attention for providing a minimalist simple Siamese (SimSiam) method to avoid collapse. However, the reason for how it avoids collapse without negative samples remains not fully clear and our investigation starts by revisiting the explanatory claims in the original SimSiam. After refuting their claims, we introduce vector decomposition for analyzing the collapse based on the gradient analysis of the l_2-normalized representation vector. This yields a unified perspective on how negative samples and SimSiam alleviate collapse. Such a unified perspective comes timely for understanding the recent progress in SSL.
****

Attention-based Interpretability with Concept Transformers
Mattia Rigotti,Christoph Miksovic,Ioana Giurgiu,Thomas Gschwind,Paolo Scotton

Attention is a mechanism that has been instrumental in driving remarkable performance gains of deep neural network models in a host of visual, NLP and multimodal tasks.
One additional notable aspect of attention is that it conveniently exposes the ``reasoning'' behind each particular output generated by the model.
Specifically, attention scores over input regions or intermediate features have been interpreted as a measure of the contribution of the attended element to the model inference.
While the debate in regard to the interpretability of attention is still not settled, researchers have pointed out the existence of architectures and scenarios that afford a meaningful interpretation of the attention mechanism.

Here we propose the generalization of attention from low-level input features to high-level concepts as a mechanism to ensure the interpretability of attention scores within a given application domain.
In particular, we design the ConceptTransformer, a deep learning module that exposes explanations of the output of a model in which it is embedded in terms of attention over user-defined high-level concepts.
Such explanations are \emph{plausible} (i.e.\ convincing to the human user) and \emph{faithful} (i.e.\ truly reflective of the reasoning process of the model).
Plausibility of such explanations is obtained by construction by training the attention heads to conform with known relations between inputs, concepts and outputs dictated by domain knowledge.
Faithfulness is achieved by design by enforcing a linear relation between the transformer value vectors that represent the concepts and their contribution to the classification log-probabilities.

We validate our ConceptTransformer module on established explainability benchmarks and show how it can be used to infuse domain knowledge into classifiers to improve accuracy, and conversely to extract concept-based explanations of classification outputs. Code to reproduce our results is available at: \url{https://github.com/ibm/concept_transformer}.
****

Inductive Relation Prediction Using Analogy Subgraph Embeddings

Jiarui Jin,Yangkun Wang,Kounianhua Du,Weinan Zhang,Zheng Zhang,David Wipf,Yong Yu,Quan Gan

Prevailing methods for relation prediction in heterogeneous graphs aim at learning latent representations (i.e., embeddings) of observed nodes and relations, and thus are limited to the transductive setting where the relation types must be known during training. Here, we propose ANalogy SubGraphEmbeddingLearning (GraphANGEL), a novel relation prediction framework that predicts relations5between each node pair based on the subgraphs containing the pair, as well as other (analogy) subgraphs with the same graph patterns.   Each graph pattern explicitly represents a specific logical rule, which contributes to an inductive bias that facilitates generalization to unseen relations and leads to more explainable predictive models. Moreover, our method also removes the limited neighborhood constraint of graph neural networks. Our model consistently outperforms existing models on heterogeneous graph based recommendation as well as knowledge graph completion. We also empirically demonstrate our model’s capability in generalizing to new relations while producing explainable heat maps of attention scores across the discovered logic.
****

High Precision Score-based Diffusion Models

Dongjun Kim,Seungjae Shin,Kyungwoo Song,Wanmo Kang,Il-chul Moon

Recent advances in diffusion models bring the state-of-the art performance on image generation tasks. However, the image generation is still an arduous task in high resolution, both theoretically and practically. From the theory side, the difficulty arises in estimating the high precision diffusion because the data score goes to ∞ as $t \rightarrow 0$ of the diffusion time. This paper resolves this difficulty by improving the previous diffusion models from three aspect

s. First, we propose an alternative parameterization for such unbounded data score, which theoretically enables the unbounded score estimation. Second, we provide a practical soft truncation method (ST-trick) to handle the extreme variation of the score scales. Third, we design a reciprocal variance exploding stochastic differential equation (RVESDE) to enable the sampling at the high precision of t. These three improvements are applicable to the variations of both NCSN and DDPM, and our improved versions are named as HNCSN and HDDPM, respectively. The experiments show that the improvements result in the state-of-the-art performances in the high resolution image generation, i.e. CelebA-HQ. Also, our ablation study empirically illustrates that all of alternative parameterization, ST-trick, and RVESDE contributes to the performance enhancement.
****

Maximum Likelihood Training of Parametrized Diffusion Model
Dongjun Kim,Byeonghu Na,Se Jung Kwon,Dongsoo Lee,Wanmo Kang,Il-chul Moon
Whereas the diverse variations of the diffusion model exist in image synthesis, the previous variations have not innovated the diffusing mechanism by maintaining the static linear diffusion. Meanwhile, it is intuitive that there would be more promising diffusion pattern adapted to the data distribution. This paper introduces such adaptive and nonlinear diffusion method for the score-based diffusion models. Unlike the static and linear VE-or-VP SDEs of the previous diffusion models, our parameterized diffusion model (PDM) learns the optimal diffusion process by combining the normalizing flow ahead of the diffusion process. Specifically, PDM utilizes the flow to non-linearly transform a data variable into a latent variable, and PDM applies the diffusion process to the transformed latent distribution with the linear diffusing mechanism. Subsequently, PDM enjoys the nonlinear and learned diffusion from the perspective of the data variable. This model structure is feasible because of the invertibility of the flow. We train PDM with the variational proxy of the log-likelihood, and we prove that the variational gap between the variational bound and the log-likelihood becomes tight when the normalizing flow becomes the optimal.
****

Reinforcement Learning in Presence of Discrete Markovian Context Evolution
Hang Ren,Aivar Sootla,Taher Jafferjee,Junxiao Shen,Jun Wang,Haitham Bou Ammar
We consider a context-dependent Reinforcement Learning (RL) setting, which is characterized by: a) an unknown finite number of not directly observable contexts; b) abrupt (discontinuous) context changes occurring during an episode; and c) Markovian context evolution. We argue that this challenging case is often met in applications and we tackle it using a Bayesian model-based approach and variational inference. We adapt a sticky Hierarchical Dirichlet Process (HDP) prior for model learning, which is arguably best-suited for infinite Markov chain modeling. We then derive a context distillation procedure, which identifies and removes spurious contexts in an unsupervised fashion. We argue that the combination of these two components allows inferring the number of contexts from data thus dealing with the context cardinality assumption. We then find the representation of the optimal policy enabling efficient policy learning using off-the-shelf RL algorithms. Finally, we demonstrate empirically (using gym environments cart-pole swing-up, drone, intersection) that our approach succeeds where state-of-the-art methods of other frameworks fail and elaborate on the reasons for such failures.
****

Optimal Transport for Long-Tailed Recognition with Learnable Cost Matrix
Hanyu Peng,Mingming Sun,Ping Li
It is attracting attention to the long-tailed recognition problem, a burning issue that has become very popular recently. Distinctive from conventional recognition is that it posits that the allocation of the training set is supremely distorted. Predictably, it will pose challenges to the generalisation behaviour of the model. Approaches to these challenges revolve into two groups: firstly, training-aware methods, with the aim of enhancing the generalisability of the model by exploiting its potential in the training period; and secondly, post-hoc correction, liberally coupled with training-aware methods, which is intended to refine the predictions to the extent possible in the post-processing stage, offering th

e advantages of simplicity and effectiveness. This paper introduces an alternative direction to do the post-hoc correction, which goes beyond the statistical methods. Mathematically, we approach this issue from the perspective of optimal transport (OT), yet, choosing the exact cost matrix when applying OT is challenging and requires expert knowledge of various tasks. To overcome this limitation, we propose to employ linear mapping to learn the cost matrix without necessary configurations adaptively. Testing our methods in practice, along with high efficiency and excellent performance, our method surpasses all previous methods and has the best performance to date.

****

A Statistical Manifold Framework for Point Cloud Data

Yonghyeon LEE,Seungyeon Kim,Jinwon Choi,Frank C. Park

A large class of problems in machine learning involve data sets in which each data point is a point cloud in $\mathbb{R}^D$. The reason that most machine learning algorithms designed for point cloud data tend to be ad hoc, and difficult to measure their performance in a uniform and quantitative way, can be traced to the lack of a rigorous mathematical characterization of this space of point cloud data. The primary contribution of this paper is a Riemannian geometric structure for point cloud data. By interpreting the point cloud data as a set of samples from some underlying probability distribution, the set of point cloud data can be given the structure of a statistical manifold, with the Fisher information metric acting as a natural Riemannian metric; this structure then leads to, e.g., distance metrics, volume forms, and other coordinate-invariant, geometrically well-defined measures needed for applications. The only requirement on the part of the user is the choice of a meaningful underlying probability distribution, which is more intuitive and natural to make than what is required in existing ad hoc formulations. Two autoencoder case studies involving point cloud data are presented to demonstrate the advantages of our statistical manifold framework: (i) interpolating between two 3D point cloud data sets to smoothly deform one object into another; (ii) transforming the latent coordinates into another with less distortion. Experiments with synthetic and large-scale standard benchmark point cloud data show more natural and intuitive shape evolutions, and improved classification accuracy for linear SVM vis-\`{a}-vis existing methods.

****

Contrastive Representation Learning for 3D Protein Structures

Pedro Hermosilla,Timo Ropinski

Learning from 3D protein structures has gained a lot of attention in the fields of protein modeling and structural bioinformatics. Unfortunately, the number of available structures is orders of magnitude lower than the number of available protein sequences. Moreover, this number is reduced even more when only annotated protein structures are considered. This makes the training of existing models difficult and prone to overfitting. To address this limitation, we introduce a new representation learning framework for 3D protein structures. Our framework uses unsupervised contrastive learning to learn meaningful representations of protein structures making use of annotated and un-annotated proteins from the Protein Data Bank. We show how these representations can be used to directly solve different tasks in the field of structural bioinformatics, such as protein function and protein structural similarity prediction. Moreover, we show how fine-tuned networks, pre-trained with our algorithm, lead to significantly improved task performance.

****

Inverse Contextual Bandits: Learning How Behavior Evolves over Time

Alihan Hüyük,Daniel Jarrett,Mihaela van der Schaar

Understanding a decision-maker's priorities by observing their behavior is critical for transparency and accountability in decision processes—such as in healthcare. Though conventional approaches to policy learning almost invariably assume stationarity in behavior, this is hardly true in practice: Medical practice is constantly evolving as clinical professionals fine-tune their knowledge over time. For instance, as the medical community's understanding of organ transplantations has progressed over the years, a pertinent question is: How have actual organ allocation policies been evolving? To give an answer, we desire a policy learning method that provides interpretable representations of decision-making, in particular capturing an agent's non-stationary knowledge of the world, as well as operating in an offline manner. First, we model the evolving behavior of decision-makers in terms of contextual bandits, and formalize the problem of Inverse Contextual Bandits ("ICB''). Second, we propose two concrete algorithms as solutions, learning parametric and non-parametric representations of an agent's behaviors. Finally, using both real and simulated data for liver transplantations, we illustrate the applicability and explainability of our method, as well as benchmarking and validating the accuracy of our algorithms.

****

PriorGrad: Improving Conditional Denoising Diffusion Models with Data-Dependent Adaptive Prior

Sang-gil Lee,Heeseung Kim,Chaehun Shin,Xu Tan,Chang Liu,Qi Meng,Tao Qin,Wei Chen,Sungroh Yoon,Tie-Yan Liu

Denoising diffusion probabilistic models have been recently proposed to generate high-quality samples by estimating the gradient of the data density. The framework assumes the prior noise as a standard Gaussian distribution, whereas the corresponding data distribution may be more complicated than the standard Gaussian distribution, which potentially introduces inefficiency in denoising the prior noise into the data sample because of the discrepancy between the data and the prior. In this paper, we propose PriorGrad to improve the efficiency of the conditional diffusion model (for example, a vocoder using a mel-spectrogram as the condition) by applying an adaptive prior derived from the data statistics based on the conditional information. We formulate the training and sampling procedures of PriorGrad and demonstrate the advantages of an adaptive prior through a theoretical analysis. Focusing on the audio domain, we consider the recently proposed diffusion-based audio generative models based on both the spectral and time domains and show that PriorGrad achieves faster convergence and superior performance, leading to an improved perceptual quality and tolerance to a smaller network capacity, and thereby demonstrating the efficiency of a data-dependent adaptive prior.

****

Target-Side Input Augmentation for Sequence to Sequence Generation

Shufang Xie,Ang Lv,Yingce Xia,Lijun Wu,Tao Qin,Tie-Yan Liu,Rui Yan

Autoregressive sequence generation, a prevalent task in machine learning and natural language processing, generates every target token conditioned on both a source input and previously generated target tokens. Previous data augmentation methods, which have been shown to be effective for the task, mainly enhance source inputs (e.g., injecting noise into the source sequence by random swapping or masking, back translation, etc.) while overlooking the target-side augmentation. In this work, we propose a target-side augmentation method for sequence generation. In training, we use the decoder output probability distributions as soft indicators, which are multiplied with target token embeddings, to build pseudo tokens. These soft pseudo tokens are then used as target tokens to enhance the training. We conduct comprehensive experiments on various sequence generation tasks, including dialog generation, machine translation, and abstractive summarization. Without using any extra labeled data or introducing additional model parameters, our method significantly outperforms strong baselines. The code is available at https://github.com/TARGET-SIDE-DATA-AUG/TSDASG.

****

Inferring Offensiveness In Images From Natural Language Supervision

Patrick Schramowski,Kristian Kersting

Probing or fine-tuning (large-scale) pre-trained models results in state-of-the-art performance for many NLP tasks and, more recently, even for computer vision tasks when combined with image data. Unfortunately, these approaches also entail severe risks. In particular, large image datasets automatically scraped from the web may contain derogatory terms as categories and offensive images, and may also underrepresent specific classes. Consequently, there is an urgent need to carefully document datasets and curate their content. Unfortunately, this process is tedious and error-prone. We show that pre-trained transformers themselves provide a methodology for the automated curation of large-scale vision datasets. Based on human-annotated examples and the implicit knowledge of a CLIP based model, we demonstrate that one can select relevant prompts for rating the offensiveness of an image.

In addition to e.g. privacy violation and pornographic content previously identified in ImageNet, we demonstrate that our approach identifies further inappropriate and potentially offensive content.

****

Lagrangian Method for Episodic Learning

Huang Bojun

This paper considers the problem of learning optimal value functions for finite-time decision tasks via saddle-point optimization of a nonlinear Lagrangian function that is derived from the Q-form Bellman optimality equation. Despite a long history of research on this topic in the literature, previous works on this general approach have been almost exclusively focusing on a linear special case known as the linear programming approach to RL/MDP. Our paper brings new perspectives to this general approach in the following aspects: 1) Inspired by the usually-used linear V-form Lagrangian, we proposed a nonlinear Q-form Lagrangian function and proved that it enjoys strong duality property in spite of its nonlinearity. The Lagrangian duality property immediately leads to a new imitation learning algorithm, which we applied to Machine Translation and obtained favorable performance on standard MT benchmark. 2) We pointed out a fundamental limit of existing works, which seeks to find minimax-type saddle points of the Lagrangian function. We proved that another class of saddle points, the maximin-type ones, turn out to have better optimality property. 3) In contrast to most previous works, our theory and algorithm are oriented to the undiscounted episode-wise reward, which is practically more relevant than the usually considered discounted-MDP setting, thus have filled a gap between theory and practice on the topic.

****

Koopman Q-learning: Offline Reinforcement Learning via Symmetries of Dynamics

Matthias Weissenbacher,Samarth Sinha,Animesh Garg,Yoshinobu Kawahara

Offline reinforcement learning leverages large datasets to train policies without interactions with the environment. The learned policies may then be deployed in real-world settings where interactions are costly or dangerous. Current algorithms over-fit to the training dataset and as a consequence perform poorly when deployed to out-of-distribution generalizations of the environment. We aim to address these limitations by learning a Koopman latent representation which allows us to infer symmetries of the system's underlying dynamic. The latter is then utilized to extend the otherwise static offline dataset during training; this constitutes a novel data augmentation framework which reflects the system's dynamic and is thus to be interpreted as an exploration of the environments phase space. To obtain the symmetries we employ Koopman theory in which nonlinear dynamics are represented in terms of a linear operator acting on the space of measurement functions of the system and thus symmetries of the dynamics may be inferred directly. We provide novel theoretical results on the existence and nature of symmetries relevant for control systems such as reinforcement learning settings. Moreover, we empirically evaluate our method on several benchmark offline reinforcement learning tasks and datasets including D4RL, Metaworld and Robosuite and find that by using our framework we consistently improve the state-of-the-art for Q-learning methods.

****

$\sbf{\delta^2}$-exploration for Reinforcement Learning

Rong Zhu,Mattia Rigotti

Effectively tackling the \emph{exploration-exploitation dilemma} is still a major challenge in reinforcement learning.

Uncertainty-based exploration strategies developed in the bandit setting could theoretically offer a principled way to trade off exploration and exploitation, but applying them to the general reinforcement learning setting is impractical due to their requirement to represent posterior distributions over models, which i s computationally intractable in generic sequential decision tasks.

Recently, \emph{Sample Average Uncertainty (SAU)} was develop as an alternative method to tackle exploration in bandit problems in a scalable way.
What makes SAU particularly efficient is that it only depends on the value predictions, meaning that it does not need to rely on maintaining model posterior distributions.
In this work we propose \emph{δ^2-exploration}, an exploration strategy that extends SAU from bandits to the general sequential Reinforcement Learning scenario.
We empirically study δ^2-exploration in the tabular as well as in the Deep Q-learning case, proving its strong practical advantage and wide adaptability to complex reward models such as those deployed in modern Reinforcement Learning.
****

UniFormer: Unified Transformer for Efficient Spatial-Temporal Representation Learning
Kunchang Li,Yali Wang,Gao Peng,Guanglu Song,Yu Liu,Hongsheng Li,Yu Qiao
It is a challenging task to learn rich and multi-scale spatiotemporal semantics from high-dimensional videos, due to large local redundancy and complex global dependency between video frames. The recent advances in this research have been mainly driven by 3D convolutional neural networks and vision transformers. Although 3D convolution can efficiently aggregate local context to suppress local redundancy from a small 3D neighborhood, it lacks the capability to capture global dependency because of the limited receptive field. Alternatively, vision transformers can effectively capture long-range dependency by self-attention mechanism, while having the limitation on reducing local redundancy with blind similarity comparison among all the tokens in each layer. Based on these observations, we propose a novel Unified transFormer (UniFormer) which seamlessly integrates merits of 3D convolution and spatiotemporal self-attention in a concise transformer format, and achieves a preferable balance between computation and accuracy. Different from traditional transformers, our relation aggregator can tackle both spatiotemporal redundancy and dependency, by learning local and global token affinity respectively in shallow and deep layers. We conduct extensive experiments on the popular video benchmarks, e.g., Kinetics-400, Kinetics-600, and Something-Something V1&V2. With only ImageNet-1K pretraining, our UniFormer achieves 82.9%/84.8% top-1 accuracy on Kinetics-400/Kinetics-600, while requiring 10x fewer GFLOPs than other state-of-the-art methods. For Something-Something V1 and V2, our UniFormer achieves new state-of-the-art performances of 60.9% and 71.2% top-1 accuracy respectively. Code is available at https://github.com/Sense-X/UniFormer.
****

EXPLAINABLE AI-BASED DYNAMIC FILTER PRUNING OF CONVOLUTIONAL NEURAL NETWORKS
Muhammad Sabih,Frank Hannig,Jürgen Teich
Filter pruning is one of the most effective ways to accelerate Convolutional Neural Networks (CNNs). Most of the existing works are focused on the static pruning of CNN filters. In dynamic pruning of CNN filters, existing works are based on the idea of switching between different branches of a CNN or exiting early based on the difficulty of a sample. These approaches can reduce the average latency of inference, but they cannot reduce the longest-path latency of inference. In contrast, we present a novel approach of dynamic filter pruning that utilizes explainable AI along with early coarse prediction in the intermediate layers of a CNN. This coarse prediction is performed using a simple branch that is trained to perform top-k classification. The branch either predicts the output class with high confidence, in which case, the rest of the computations are left out. Alternatively, the branch predicts the output class to be within a subset of possible output classes. After this coarse prediction, only those filters that are important for this subset of classes are utilized for further computations. The importances of filters for each output class are obtained using explainable AI. Using this architecture of dynamic pruning, we not only reduce the average latency of inference, but we can also reduce the longest-path latency of inference. Our pr

oposed architecture for dynamic pruning can be deployed on different hardware platforms. We evaluate our approach using commonly used image classification models and datasets on CPU and GPU platforms and demonstrate speedup without significant overhead.

****

ClsVC: Learning Speech Representations with two different classification tasks.
Tang huaizhen,xulong Zhang,Jianzong Wang,ning Cheng,Jing Xiao

Voice conversion(VC) aims to convert one speaker's voice to generate a new speech as it is said by another speaker. Previous works focus on learning latent representation by applying two different encoders to learn content information and timbre information from the input speech respectively. However, whether they apply a bottleneck network or vector quantify technology, it is very difficult to perfectly separate the speaker and the content information from a speech signal. In this paper, we propose a novel voice conversion framework, 'ClsVC', to address this problem. It uses only one encoder to get both timbre and content information by dividing the latent space. Besides, some constraints are proposed to ensure the different part of latent space only contains separating content and timbre information respectively. We have shown the necessity to set these constraints, and we also experimentally prove that even if we change the division proportion of latent space, the content and timbre information will be always well separated. Experiments on the VCTK dataset show ClsVC is a state-of-the-art framework in terms of the naturalness and similarity of converted speech.

****

Inverse Online Learning: Understanding Non-Stationary and Reactionary Policies
Alex Chan,Alicia Curth,Mihaela van der Schaar

Human decision making is well known to be imperfect and the ability to analyse such processes individually is crucial when attempting to aid or improve a decision-maker's ability to perform a task, e.g. to alert them to potential biases or oversights on their part. To do so, it is necessary to develop interpretable representations of how agents make decisions and how this process changes over time as the agent learns online in reaction to the accrued experience. To then understand the decision-making processes underlying a set of observed trajectories, we cast the policy inference problem as the inverse to this online learning problem. By interpreting actions within a potential outcomes framework, we introduce a meaningful mapping based on agents choosing an action they believe to have the greatest treatment effect. We introduce a practical algorithm for retrospectively estimating such perceived effects, alongside the process through which agents update them, using a novel architecture built upon an expressive family of deep state-space models. Through application to the analysis of UNOS organ donation acceptance decisions, we demonstrate that our approach can bring valuable insights into the factors that govern decision processes and how they change over time.

****

Multi-Mode Deep Matrix and Tensor Factorization
Jicong Fan

Recently, deep linear and nonlinear matrix factorizations gain increasing attention in the area of machine learning. Existing deep nonlinear matrix factorization methods can only exploit partial nonlinearity of the data and are not effective in handling matrices of which the number of rows is comparable to the number of columns. On the other hand, there is still a gap between deep learning and tensor decomposition. This paper presents a framework of multi-mode deep matrix and tensor factorizations to explore and exploit the full nonlinearity of the data in matrices and tensors. We use the factorization methods to solve matrix and tensor completion problems and prove that our methods have tighter generalization error bounds than conventional matrix and tensor factorization methods. The experiments on synthetic data and real datasets showed that the proposed methods have much higher recovery accuracy than many baselines.

****

LORD: Lower-Dimensional Embedding of Log-Signature in Neural Rough Differential Equations

JAEHOON LEE,Jinsung Jeon,Sheo yon Jhin,Jihyeon Hyeong,Jayoung Kim,Minju Jo,Kook Seungji,Noseong Park

The problem of processing very long time-series data (e.g., a length of more than 10,000) is a long-standing research problem in machine learning. Recently, one breakthrough, called neural rough differential equations (NRDEs), has been proposed and has shown that it is able to process such data. Their main concept is to use the log-signature transform, which is known to be more efficient than the Fourier transform for irregular long time-series, to convert a very long time-series sample into a relatively shorter series of feature vectors. However, the log-signature transform causes non-trivial spatial overheads. To this end, we present the method of LOweR-Dimensional embedding of log-signature (LORD), where we define an NRDE-based autoencoder to implant the higher-depth log-signature knowledge into the lower-depth log-signature. We show that the encoder successfully combines the higher-depth and the lower-depth log-signature knowledge, which greatly stabilizes the training process and increases the model accuracy. In our experiments with benchmark datasets, the improvement ratio by our method is up to 75\% in terms of various classification and forecasting evaluation metrics.

****

Divergence-Regularized Multi-Agent Actor-Critic

Su Kefan,Zongqing Lu

Entropy regularization is a popular method in reinforcement learning (RL). Although it has many advantages, it alters the RL objective and makes the converged policy deviate from the optimal policy of the original Markov Decision Process. Though divergence regularization has been proposed to settle this problem, it cannot be trivially applied to cooperative multi-agent reinforcement learning (MARL). In this paper, we investigate divergence regularization in cooperative MARL and propose a novel off-policy cooperative MARL framework, divergence-regularized multi-agent actor-critic (DMAC). Mathematically, we derive the update rule of DMAC which is naturally off-policy, guarantees a monotonic policy improvement and is not biased by the regularization. DMAC is a flexible framework and can be combined with many existing MARL algorithms. We evaluate DMAC in a didactic stochastic game and StarCraft Multi-Agent Challenge and empirically show that DMAC substantially improves the performance of existing MARL algorithms.

****

Network Pruning Optimization by Simulated Annealing Algorithm

Chun Lin Kuo,Ercan Engin Kuruoglu,Wai Kin Victor Chan

One critical problem of large neural networks is over-parameterization with a large number of weight parameters. This becomes an obstacle to implement networks in edge devices as well as limiting the development of industrial applications by engineers for machine learning problems. Plenty of papers have shown that the redundant branches can be erased strategically in a fully connected network. In this work, we reduce network complexity by pruning and structure optimization. We propose to do network optimization by Simulated Annealing, a heuristic based non-convex optimization method which can potentially solve this NP-hard problem and find the global minimum for a given percentage of branch pruning given sufficient amount of time. Our results have shown that Simulated Annealing can significantly reduce the complexity of a fully connected neural network with only limited loss of performance.

****

Generalized Natural Gradient Flows in Hidden Convex-Concave Games and GANs

Andjela Mladenovic,Iosif Sakos,Gauthier Gidel,Georgios Piliouras

Game-theoretic formulations in machine learning have recently risen in prominence, whereby entire modeling paradigms are best captured as zero-sum games. Despite their popularity, however, their dynamics are still poorly understood. This lack of theory is often substantiated with painful empirical observations of volatile training dynamics and even divergence. Such results highlight the need to develop an appropriate theory with convergence guarantees that are powerful enough to inform practice. This paper studies the generalized Gradient Descent-Ascent (GDA) flow in a large class of non-convex non-concave Zero-Sum games dubbed Hidden Convex-Concave games, a class of games that includes GANs. We focus on two sp

ecific geometries: a novel geometry induced by the hidden convex-concave structure that we call the hidden mapping geometry and the Fisher information geometry. For the hidden mapping geometry, we prove global convergence under mild assumptions. In the case of Fisher information geometry, we provide a complete picture of the dynamics in an interesting special setting of team competition via invariant function analysis.
****

Bootstrapped Hindsight Experience replay with Counterintuitive Prioritization
Jiawei Xu,Shuxing Li,Chun Yuan,Zhengyou Zhang,Lei Han
Goal-conditioned environments are known as sparse rewards tasks, in which the agent gains a positive reward only when it achieves the goal. Such an setting results in much difficulty for the agent to explore successful trajectories. Hindsight experience replay (HER) replaces the goal in failed experiences with any practically achieved one, so that the agent has a much higher chance to see successful trajectories even if they are fake. Comprehensive results have demonstrated the effectiveness of HER in the literature. However, the importance of the fake trajectories differs in terms of exploration and exploitation, and it is usually inefficient to learn with a fixed proportion of fake and original data as HER did. In this paper, inspired by Bootstrapped DQN, we use multiple heads in DDPG and take advantage of the diversity and uncertainty among multiple heads to improve the data efficiency with relabeled goals. The method is referred to as Bootstrapped HER (BHER). Specifically, in addition to the benefit from the Bootstrapped version, we explicitly leverage the uncertainty measured by the variance of estimated Q-values from multiple heads. A common knowledge is that higher uncertainty will promote exploration and hence maximizing the uncertainty via a bonus term will induce better performance in Q-learning. However, in this paper, we reveal a counterintuitive conclusion that for hindsight experiences, exploiting lower uncertainty data samples will significantly improve the performance. The explanation behind this fact is that hindsight relabeling largely promotes exploration, and then exploiting lower uncertainty data (whose goals are generated by hindsight relabeling) provides a good trade-off between exploration and exploitation, resulting in further improved data efficiency. Comprehensive experiments demonstrate that our method can achieve state-of-the-art results in many goal-conditioned tasks.
****

Offline Neural Contextual Bandits: Pessimism, Optimization and Generalization
Thanh Nguyen-Tang,Sunil Gupta,A. Tuan Nguyen,Svetha Venkatesh
Offline policy learning (OPL) leverages existing data collected a priori for policy optimization without any active exploration. Despite the prevalence and recent interest in this problem, its theoretical and algorithmic foundations in function approximation settings remain under-developed. In this paper, we consider this problem on the axes of distributional shift, optimization, and generalization in offline contextual bandits with neural networks. In particular, we propose a provably efficient offline contextual bandit with neural network function approximation that does not require any functional assumption on the reward. We show that our method provably generalizes over unseen contexts under a milder condition for distributional shift than the existing OPL works. Notably, unlike any other OPL method, our method learns from the offline data in an online manner using stochastic gradient descent, allowing us to leverage the benefits of online learning into an offline setting. Moreover, we show that our method is more computationally efficient and has a better dependence on the effective dimension of the neural network than an online counterpart. Finally, we demonstrate the empirical effectiveness of our method in a range of synthetic and real-world OPL problems.
****

THOMAS: Trajectory Heatmap Output with learned Multi-Agent Sampling
Thomas Gilles,Stefano Sabatini,Dzmitry Tsishkou,Bogdan Stanciulescu,Fabien Moutarde
In this paper, we propose THOMAS, a joint multi-agent trajectory prediction framework allowing for an efficient and consistent prediction of multi-agent multi-m

odal trajectories. We present a unified model architecture for simultaneous agent future heatmap estimation, in which we leverage hierarchical and sparse image generation for fast and memory-efficient inference. We propose a learnable trajectory recombination model that takes as input a set of predicted trajectories for each agent and outputs its consistent reordered recombination. This recombination module is able to realign the initially independent modalities so that they do no collide and are coherent with each other. We report our results on the Interaction multi-agent prediction challenge and rank 1^{st} on the online test leaderboard.

****

Correct-N-Contrast: a Contrastive Approach for Improving Robustness to Spurious Correlations

Michael Zhang,Nimit Sharad Sohoni,Hongyang Zhang,Chelsea Finn,Christopher Re

Spurious correlations pose a fundamental challenge for building robust machine learning models. For example, models trained with empirical risk minimization (ERM) may depend on correlations between class labels and spurious features to classify data, even if these relations only hold for certain data groups. This can result in poor performance on other groups that do not exhibit such relations. When group information is available during training, Sagawa et al. (2019) have shown how to improve worst-group performance by optimizing the worst-group loss (GDRO). However, when group information is unavailable, improving worst-group performance is more challenging. For this latter setting, we propose Correct-N-Contrast (CNC), a contrastive learning method to train models more robust to spurious correlations. Our motivating observation is that worst-group performance is related to a representation alignment loss, which measures the distance in feature space between different groups within each class. We prove that the gap between worst-group and average loss for each class is upper bounded by the alignment loss for that class. Thus, CNC aims to improve representation alignment via contrastive learning. First, CNC uses an ERM model to infer the group information. Second, with a careful sampling scheme, CNC trains a contrastive model to encourage similar representations for groups in the same class. We show that CNC significantly improves worst-group accuracy over existing state-of-the-art methods on popular benchmarks, e.g., achieving 7.7% absolute lift in worst-group accuracy on the CelebA data set, and performs almost as well as GDRO trained with group labels. CNC also learns better-aligned representations between different groups in each class, reducing the alignment loss substantially compared to prior methods.

****

Robust Imitation via Mirror Descent Inverse Reinforcement Learning

Dong-Sig Han,Hyunseo Kim,Hyundo Lee,JeHwan Ryu,Byoung-Tak Zhang

Adversarial imitation learning techniques are based on modeling statistical divergences using agent and expert demonstration data. However, unbiased minimization of these divergences is not usually guaranteed due to the geometry of the underlying space. Furthermore, when the size of demonstrations is not sufficient, estimated reward functions from the discriminative signals become uncertain and fail to give informative feedback. Instead of formulating a global cost at once, we consider reward functions as an iterative sequence in a proximal method. In this paper, we show that rewards dervied by mirror descent ensures minimization of a Bregman divergence in terms of a rigorous regret bound of $\mathcal{O}(1/T)$ for a particular condition of step sizes $\{\eta_t\}_{t=1}^T$. The resulting mirror descent adversarial inverse reinforcement learning (MD-AIRL) algorithm gradually advances a parameterized reward function in an associated reward space, and the sequence of such functions provides optimization targets for the policy space. We empirically validate our method in discrete and continuous benchmarks and show that MD-AIRL outperforms previous methods in various settings.

****

Fast Deterministic Stackelberg Actor-Critic

Runsheng Yu,Xinrun Wang,James Kwok

Most advanced Actor-Critic (AC) approaches update the actor and critic concurrently through (stochastic) Gradient Descents (GD), which may be trapped into bad local optimality due to the instability of these simultaneous updating schemes.

Stackelberg AC learning scheme alleviates these limitations by adding a compensated indirect gradient terms to the GD. However, the indirect gradient terms are time-consuming to calculate, and the convergence rate is also relatively slow. To alleviates these challenges, we find that in the Deterministic Policy Gradient family, by removing the terms that contain Hessian matrices and adopting the block diagonal approximation technique to approximate the remaining inverse matrices, we can construct an approximated Stackelberg AC learning scheme that is easy to compute and fast to converge. Experiments reveal that ours outperform SOTAs in terms of average returns under acceptable training time.
****

A Free Lunch from the Noise: Provable and Practical Exploration for Representation Learning

Tongzheng Ren,Tianjun Zhang,Csaba Szepesvari,Bo Dai

Representation learning lies at the heart of the empirical success of deep learning for dealing with the curse of dimensionality. However, the power of representation learning has not been fully exploited yet in reinforcement learning (RL), due to i), the trade-off between expressiveness and tractability; and ii), the coupling between exploration and representation learning. In this paper, we first reveal the fact that under some noise assumption in the stochastic control model, we can obtain the linear spectral feature of its corresponding Markov transition operator in closed-form for free. Based on this observation, we propose Spectral Dynamics Embedding (SPEDE), which breaks the trade-off and completes optimistic exploration for representation learning by exploiting the structure of the noise. We provide rigorous theoretical analysis of SPEDE, and demonstrate the practical superior performance over the existing state-of-the-art empirical algorithms on several benchmarks.

****

CLEVA-Compass: A Continual Learning Evaluation Assessment Compass to Promote Research Transparency and Comparability

Martin Mundt,Steven Lang,Quentin Delfosse,Kristian Kersting

What is the state of the art in continual machine learning? Although a natural question for predominant static benchmarks, the notion to train systems in a lifelong manner entails a plethora of additional challenges with respect to set-up and evaluation. The latter have recently sparked a growing amount of critiques on prominent algorithm-centric perspectives and evaluation protocols being too narrow, resulting in several attempts at constructing guidelines in favor of specific desiderata or arguing against the validity of prevalent assumptions. In this work, we depart from this mindset and argue that the goal of a precise formulation of desiderata is an ill-posed one, as diverse applications may always warrant distinct scenarios. Instead, we introduce the Continual Learning EValuation Assessment Compass: the CLEVA-Compass. The compass provides the visual means to both identify how approaches are practically reported and how works can simultaneously be contextualized in the broader literature landscape. In addition to promoting compact specification in the spirit of recent replication trends, it thus provides an intuitive chart to understand the priorities of individual systems, where they resemble each other, and what elements are missing towards a fair comparison.
****

Equal Experience in Recommender Systems

Jaewoong Cho,Moonseok Choi,Changho Suh

We explore the fairness issue that arises in recommender systems. Biased data due to inherent stereotypes of particular groups (e.g., male students' average rating on mathematics is often higher than that on humanities, and vice versa for females) may yield a limited scope of suggested items to a certain group of users. Our main contribution lies in the introduction of a novel fairness notion (that we call equal experience), which can serve to regulate such unfairness in the presence of biased data. The notion captures the degree of the equal experience of item recommendations across distinct groups. We propose an optimization framework that incorporates the fairness notion as a regularization term, as well as introduce computationally-efficient algorithms that solve the optimization. Experiments on synthetic and benchmark real datasets demonstrate that the proposed framework can indeed mitigate such unfairness while exhibiting a minor degradation of recommendation accuracy.

****

Generating Novel Scene Compositions from Single Images and Videos

Vadim Sushko,Dan Zhang,Juergen Gall,Anna Khoreva

Given a large dataset for training, GANs can achieve remarkable performance for the image synthesis task. However, training GANs in extremely low data regimes remains a challenge, as overfitting often occurs, leading to memorization or training divergence. In this work, we introduce SIV-GAN, an unconditional generative model that can generate new scene compositions from a single training image or a single video clip. We propose a two-branch discriminator architecture, with content and layout branches designed to judge internal content and scene layout realism separately from each other. This discriminator design enables synthesis of visually plausible, novel compositions of a scene, with varying content and layout, while preserving the context of the original sample. Compared to previous single image GANs, our model generates more diverse, higher quality images, while not being restricted to a single image setting. We show that SIV-GAN successfully deals with a new challenging task of learning from a single video, for which prior GAN models fail to achieve synthesis of both high quality and diversity.

****

Neural Stochastic Dual Dynamic Programming

Hanjun Dai,Yuan Xue,Zia Syed,Dale Schuurmans,Bo Dai

Stochastic dual dynamic programming (SDDP) is a state-of-the-art method for solving multi-stage stochastic optimization, widely used for modeling real-world process optimization tasks. Unfortunately, SDDP has a worst-case complexity that scales exponentially in the number of decision variables, which severely limits applicability to only low dimensional problems. To overcome this limitation, we extend SDDP by introducing a trainable neural model that learns to map problem instances to a piece-wise linear value function within intrinsic low-dimension space, which is architected specifically to interact with a base SDDP solver, so that can accelerate optimization performance on new instances. The proposed Neural Stochastic Dual Dynamic Programming ($$\nu$$-SDDP) continually self-improves by solving successive problems. An empirical investigation demonstrates that $$\nu$$-SDDP can significantly reduce problem solving cost without sacrificing solution quality over competitors such as SDDP and reinforcement learning algorithms, across a range of synthetic and real-world process optimization problems.

****

Off-Policy Reinforcement Learning with Delayed Rewards

Beining Han,Zhizhou Ren,Zuofan Wu,Yuan Zhou,Jian Peng

We study deep reinforcement learning (RL) algorithms with delayed rewards. In many real-world tasks, instant rewards are often not readily accessible or even defined immediately after the agent performs actions. In this work, we first formally define the environment with delayed rewards and discuss the challenges raised due to the non-Markovian nature of such environments. Then, we introduce a general off-policy RL framework with a new Q-function formulation that can handle the delayed rewards with theoretical convergence guarantees. For practical tasks with high dimensional state spaces, we further introduce the HC-decomposition rule of the Q-function in our framework which naturally leads to an approximation scheme that helps boost the training efficiency and stability. We finally conduct extensive experiments to demonstrate the superior performance of our algorithms over the existing work and their variants.

****

Learning Multimodal VAEs through Mutual Supervision

Tom Joy,Yuge Shi,Philip Torr,Tom Rainforth,Sebastian M Schmon,Siddharth N

Multimodal VAEs seek to model the joint distribution over heterogeneous data (e.g.\ vision, language), whilst also capturing a shared representation across such modalities. Prior work has typically combined information from the modalities by reconciling idiosyncratic representations directly in the recognition model through explicit products, mixtures, or other such factorisations. Here we introduce a novel alternative, the MEME, that avoids such explicit combinations by repurposing semi-supervised VAEs to combine information between modalities implicitly through mutual supervision. This formulation naturally allows learning from partially-observed data where some modalities can be entirely missing---something that most existing approaches either cannot handle, or do so to a limited extent. We demonstrate that MEME outperforms baselines on standard metrics across both partial and complete observation schemes on the MNIST-SVHN (image--image) and CUB (image--text) datasets. We also contrast the quality of the representations learnt by mutual supervision against standard approaches and observe interesting trends in its ability to capture relatedness between data.

****

DemoDICE: Offline Imitation Learning with Supplementary Imperfect Demonstrations

Geon-Hyeong Kim,Seokin Seo,Jongmin Lee,Wonseok Jeon,HyeongJoo Hwang,Hongseok Yang,Kee-Eung Kim

We consider offline imitation learning (IL), which aims to mimic the expert's behavior from its demonstration without further interaction with the environment. One of the main challenges in offline IL is to deal with the narrow support of the data distribution exhibited by the expert demonstrations that cover only a small fraction of the state and the action spaces. As a result, offline IL algorithms that rely only on expert demonstrations are very unstable since the situation easily deviates from those in the expert demonstrations. In this paper, we assume additional demonstration data of unknown degrees of optimality, which we call imperfect demonstrations. Under this setting, we propose DemoDICE, which effectively utilizes imperfect demonstrations by matching the stationary distribution of a policy with experts' distribution while penalizing its deviation from the overall demonstrations. Compared with the recent IL algorithms that adopt adversarial minimax training objectives, we substantially stabilize overall learning process by reducing minimax optimization to a direct convex optimization in a principled manner. Using extensive tasks, we show that DemoDICE achieves promising results in the offline IL from expert and imperfect demonstrations.

****

When should agents explore?

Miruna Pislar,David Szepesvari,Georg Ostrovski,Diana L Borsa,Tom Schaul

Exploration remains a central challenge for reinforcement learning (RL). Virtually all existing methods share the feature of a *monolithic* behaviour policy that changes only gradually (at best). In contrast, the exploratory behaviours of animals and humans exhibit a rich diversity, namely including forms of *switching* between modes. This paper presents an initial study of mode-switching, non-monolithic exploration for RL. We investigate different modes to switch between, at what timescales it makes sense to switch, and what signals make for good switching triggers. We also propose practical algorithmic components that make the switching mechanism adaptive and robust, which enables flexibility without an accompanying hyper-parameter-tuning burden. Finally, we report a promising initial study on Atari, using two-mode exploration and switching at sub-episodic time-scales.

****

Adversarial robustness against multiple l_p-threat models at the price of one and how to quickly fine-tune robust models to another threat model

Francesco Croce,Matthias Hein

Adversarial training (AT) in order to achieve adversarial robustness wrt single l_p-threat models has been discussed extensively. However, for safety-critical systems adversarial robustness should be achieved wrt all l_p-threat models simultaneously. In this paper we develop a simple and efficient training scheme to achieve adversarial robustness against the union of l_p-threat models. Our novel E-AT scheme is based on geometric considerations of the different l_p-balls and costs as much as normal adversarial training against a single l_p-threat model. Moreover, we show that using our E-AT scheme one can fine-tune with jus

t 3 epochs \emph{any} l_p-robust model (for $p \in \{1,2,\infty\}$) and achieve multiple norm adversarial robustness. In this way we boost the state-of-the-art for multiple-norm robustness to more than 51% on CIFAR-10 and report up to our knowledge the first ImageNet models with multiple norm robustness. Moreover, we study the general transfer of adversarial robustness between different threat models and in this way boost the previous SOTA l_1-robustness on CIFAR-10 by almost 10%.
****

Distributed Skellam Mechanism: a Novel Approach to Federated Learning with Differential Privacy

Ergute Bao,Yizheng Zhu,Xiaokui Xiao,Yin Yang,Beng Chin Ooi,Benjamin Hong Meng Tan,Khin Mi Mi Aung

Deep neural networks have strong capabilities of memorizing the underlying training data; on the flip side, unintended data memorization can be a serious privacy concern. An effective and rigorous approach to addressing this problem is to train models with \textit{differential privacy} (\textit{DP}), which provides information-theoretic privacy guarantees by injecting random noise to the gradients. This paper focuses on the scenario where sensitive data are distributed among individual participants, who jointly train a model through \textit{federated learning}, using both \textit{secure multiparty computation} (\textit{MPC}) to ensure the confidentiality of individual gradient updates, and differential privacy to avoid data leakage in the resulting model. We point out a major challenge in this problem setting: that common mechanisms for enforcing DP in deep learning, which require injecting \textit{real-valued noise}, are fundamentally incompatible with MPC, which exchanges \textit{finite-field integers} among the participants. Consequently, existing DP mechanisms require rather high noise levels, leading to poor model utility.

Motivated by this, we design and develop \textit{distributed Skellam mechanism} ({\sf DSM}), a novel solution for enforcing differential privacy on models built through an MPC-based federated learning process. Compared to existing approaches, {\sf DSM} has the advantage that its privacy guarantee is independent of the dimensionality of the gradients; further, {\sf DSM} allows tight privacy accounting due to the nice composition and sub-sampling properties of the Skellam distribution, which are key to accurate deep learning with DP. The theoretical analysis of {\sf DSM} is highly non-trivial, especially considering (i) the complicated math of differentially private deep learning in general and (ii) the fact that the Skellam distribution is rather complex, and to our knowledge, has not been applied to an iterative and sampling-based process, i.e., stochastic gradient descent. Meanwhile, through extensive experiments on various practical settings, we demonstrate that {\sf DSM} consistently outperforms existing solutions in terms of model utility by a large margin.
****

Learning to Extend Molecular Scaffolds with Structural Motifs

Krzysztof Maziarz,Henry Richard Jackson-Flux,Pashmina Cameron,Finton Sirockin,Nadine Schneider,Nikolaus Stiefl,Marwin Segler,Marc Brockschmidt

Recent advancements in deep learning-based modeling of molecules promise to accelerate in silico drug discovery. A plethora of generative models is available, building molecules either atom-by-atom and bond-by-bond or fragment-by-fragment. However, many drug discovery projects require a fixed scaffold to be present in the generated molecule, and incorporating that constraint has only recently been explored. Here, we propose MoLeR, a graph-based model that naturally supports scaffolds as initial seed of the generative procedure, which is possible because it is not conditioned on the generation history. Our experiments show that MoLeR performs comparably to state-of-the-art methods on unconstrained molecular optimization tasks, and outperforms them on scaffold-based tasks, while being an order of magnitude faster to train and sample from than existing approaches. Furthermore, we show the influence of a number of seemingly minor design choices on the overall performance.
****

Adaptive Label Smoothing with Self-Knowledge

Dongkyu Lee,Ka Chun Cheung,Nevin Zhang

Overconfidence has been shown to impair generalization and calibration of a neural network. Previous studies remedy this issue by adding a regularization term to a loss function, preventing a model from making a peaked distribution. Label smoothing smoothes target labels with a predefined prior label distribution; as a result, a model is learned to maximize the likelihood of predicting the soft label. Nonetheless, the amount of smoothing is the same in all samples and remains fixed in training. In other words, label smoothing does not reflect the change in probability distribution mapped by a model over the course of training. To address this issue, we propose a regularization scheme that brings dynamic nature into the smoothing parameter by taking model probability distribution into account, thereby varying the parameter per instance. A model in training self-regulates the extent of smoothing on the fly during forward propagation. Furthermore, inspired by recent work in bridging label smoothing and knowledge distillation, our work utilizes self-knowledge as a prior label distribution in softening target labels, and presents theoretical support for the regularization effect by knowledge distillation. Our regularizer is validated comprehensively on various datasets in machine translation and outperforms strong baselines not only in model performance but also in model calibration by a large margin.

****

Do Androids Dream of Electric Fences? Safety-Aware Reinforcement Learning with Latent Shielding

Peter He,Borja G. León,Francesco Belardinelli

The growing trend of fledgling reinforcement learning systems making their way into real-world applications has been accompanied by growing concerns for their safety and robustness. In recent years, a variety of approaches have been put forward to address the challenges of safety-aware reinforcement learning; however, these methods often either require a handcrafted model of the environment to be provided beforehand, or that the environment is relatively simple and low-dimensional. We present a novel approach to safety-aware deep reinforcement learning in high-dimensional environments called latent shielding. Latent shielding leverages internal representations of the environment learnt by model-based agents to "imagine" future trajectories and avoid those deemed unsafe. We experimentally demonstrate that this approach leads to improved adherence to formally-defined safety specifications.

****

Sharp Learning Bounds for Contrastive Unsupervised Representation Learning

Han Bao,Yoshihiro Nagano,Kento Nozawa

Contrastive unsupervised representation learning (CURL) encourages data representation to make semantically similar pairs closer than randomly drawn negative samples, which has been successful in various domains such as vision, language, and graphs. Although recent theoretical studies have attempted to explain its success by upper bounds of a downstream classification loss by the contrastive loss, they are still not tight enough to explain an experimental fact: larger negative samples improve the classification performance. This study establishes a downstream classification loss bound with a tight intercept in the negative sample size. By regarding the contrastive loss as a downstream loss estimator, our theory not only improves the existing learning bounds substantially but also explains why downstream classification empirically improves with larger negative samples because the estimation variance of the downstream loss decays with larger negative samples. We verify that our theory is consistent with experiments on synthetic, vision, and language datasets.

****

Understanding the robustness-accuracy tradeoff by rethinking robust fairness

Zihui Wu,Haichang Gao,Shudong Zhang,Yipeng Gao

Although current adversarial training (AT) methods can effectively improve the robustness on adversarial examples,

they usually lead to a decrease in accuracy, called the robustness-accuracy trade-off. In addition, researchers have recently discovered a robust fairness pheno menon in the AT model; that is, not all categories of the dataset have experienced a serious decline in accuracy with the introduction of AT methods. In this paper, we explore the relationship between the robustness-accuracy tradeoff and robust fairness for the first time. Empirically, we have found that AT will cause a substantial increase in the inter-class similarity, which could be the root cause of these two phenomena. We argue that the label smoothing (LS) is more than a trick in AT. The smoothness learned from LS can help reduce the excessive inter-class similarity caused by AT, and also reduce the intra-class variance, thereby significantly improving accuracy. Then, we explored the effect of another classic smoothing regularizer, namely, the maximum entropy (ME), and we have found ME can also help reduce both inter-class similarity and intra-class variance. Additionally, we revealed that TRADES actually implies the function of ME, which can explain why TRADES usually performs better than PGD-AT on robustness. Finally, we proposed the maximum entropy PGD-AT (ME-AT) and the maximum entropy TRADES (ME-TRADES), and experimental results show that our methods can significantly mitigate both tradeoff and robust fairness.

****

Diverse and Consistent Multi-view Networks for Semi-supervised Regression

Cuong Manh Nguyen,Le Zhang,Arun Raja,Xun Xu,Balagopal Unnikrishnan,Kangkang Lu,Chuan-Sheng Foo

Label collection is costly in many applications, which poses the need for label-efficient learning. In this work, we present Diverse and Consistent Multi-view Networks (DiCoM) — a novel semi-supervised regression technique based on a multi-view learning framework. DiCoM combines diversity with consistency — two seemingly opposing yet complementary principles of multi-view learning - based on underlying probabilistic graphical assumptions. Given multiple deep views of the same input, DiCoM encourages a negative correlation among the views' predictions on labeled data, while simultaneously enforces their agreement on unlabeled data. DiCoM can utilize either multi-network or multi-branch architectures to make a trade-off between computational cost and modeling performance. Under realistic evaluation setups, DiCoM outperforms competing methods on tabular and image data. Our ablation studies confirm the importance of having both consistency and diversity.

****

Revisiting Layer-wise Sampling in Fast Training for Graph Convolutional Networks

Yifan Chen,Tianning Xu,Dilek Hakkani-Tur,Di Jin,Yun Yang,Ruoqing Zhu

To accelerate the training of graph convolutional networks (GCN), many sampling-based methods have been developed for approximating the embedding aggregation. Among them, a layer-wise approach recursively performs importance sampling to select neighbors jointly for existing nodes in each layer. This paper revisits the approach from a matrix approximation perspective. We identify two issues in the existing layer-wise sampling methods: sub-optimal sampling probabilities and the approximation bias induced by sampling without replacement. We thus propose remedies to address these issues. The improvements are demonstrated by extensive analyses and experiments on common benchmarks.

****

Discrepancy-Based Active Learning for Domain Adaptation

Antoine de Mathelin,François Deheeger,Mathilde MOUGEOT,Nicolas Vayatis

The goal of the paper is to design active learning strategies which lead to domain adaptation under an assumption of Lipschitz functions. Building on previous work by Mansour et al. (2009) we adapt the concept of discrepancy distance between source and target distributions to restrict the maximization over the hypothesis class to a localized class of functions which are performing accurate labeling on the source domain. We derive generalization error bounds for such active learning strategies in terms of Rademacher average and localized discrepancy for general loss functions which satisfy a regularity condition. A practical K-medoids algorithm that can address the case of large data set is inferred from the theoretical bounds. Our numerical experiments show that the proposed algorithm is competitive against other state-of-the-art active learning techniques in the context of domain adaptation, in particular on large data sets of around one hundred thousand images.

****

Gradient Matching for Domain Generalization

Yuge Shi,Jeffrey Seely,Philip Torr,Siddharth N,Awni Hannun,Nicolas Usunier,Gabriel Synnaeve

Machine learning systems typically assume that the distributions of training and test sets match closely. However, a critical requirement of such systems in the real world is their ability to generalize to unseen domains. Here, we propose an _inter-domain gradient matching_ objective that targets domain generalization by maximizing the inner product between gradients from different domains. Since direct optimization of the gradient inner product can be computationally prohibitive --- it requires computation of second-order derivatives –-- we derive a simpler first-order algorithm named Fish that approximates its optimization. We perform experiments on the Wilds benchmark, which captures distribution shift in the real world, as well as the DomainBed benchmark that focuses more on synthetic-to-real transfer. Our method produces competitive results on both benchmarks, demonstrating its effectiveness across a wide range of domain generalization tasks.

****

Knowledge Based Multilingual Language Model

Linlin Liu,Xin Li,Ruidan He,Lidong Bing,Shafiq Joty,Luo Si

Knowledge enriched language representation learning has shown promising performance across various knowledge-intensive NLP tasks. However, existing knowledge based language models are all trained with monolingual knowledge graph data, which limits their application to more languages. In this work, we present a novel framework to pretrain knowledge based multilingual language models (KMLMs). We first generate a large amount of code-switched synthetic sentences and reasoning-based multilingual training data using the Wikidata knowledge graphs. Then based on the intra- and inter-sentence structures of the generated data, we design pretraining tasks to facilitate knowledge learning, which allows the language models to not only memorize the factual knowledge but also learn useful logical patterns. Our pretrained KMLMs demonstrate significant performance improvements on a wide range of knowledge-intensive cross-lingual NLP tasks, including named entity recognition, factual knowledge retrieval, relation classification, and a new task designed by us, namely, logic reasoning. Our code and pretrained language models will be made publicly available.

****

Revisiting Design Choices in Offline Model Based Reinforcement Learning

Cong Lu,Philip Ball,Jack Parker-Holder,Michael Osborne,Stephen J. Roberts

Offline reinforcement learning enables agents to leverage large pre-collected datasets of environment transitions to learn control policies, circumventing the need for potentially expensive or unsafe online data collection. Significant progress has been made recently in offline model-based reinforcement learning, approaches which leverage a learned dynamics model. This typically involves constructing a probabilistic model, and using the model uncertainty to penalize rewards where there is insufficient data, solving for a pessimistic MDP that lower bounds the true MDP. Existing methods, however, exhibit a breakdown between theory and practice, whereby pessimistic return ought to be bounded by the total variation distance of the model from the true dynamics, but is instead implemented through a penalty based on estimated model uncertainty. This has spawned a variety of uncertainty heuristics, with little to no comparison between differing approaches. In this paper, we compare these heuristics, and design novel protocols to investigate their interaction with other hyperparameters, such as the number of models, or imaginary rollout horizon. Using these insights, we show that selecting these key hyperparameters using Bayesian Optimization produces superior configurations that are vastly different to those currently used in existing hand-tuned state-of-the-art methods, and result in drastically stronger performance.

****

NASPY: Automated Extraction of Automated Machine Learning Models

Xiaoxuan Lou,Shangwei Guo,Jiwei Li,Yaoxin Wu,Tianwei Zhang

We present NASPY, an end-to-end adversarial framework to extract the networkarchitecture of deep learning models from Neural Architecture Search (NAS). Existing works about model extraction attacks mainly focus on conventional DNN models with very simple operations, or require heavy manual analysis with lots of domain knowledge. In contrast, NASPY introduces seq2seq models to automatically identify novel and complicated operations (e.g., separable convolution,dilated convolution) from hardware side-channel sequences. We design two models (RNN-CTC and transformer), which can achieve only 3.2% and 11.3% error rates for operation prediction. We further present methods to recover the model hyper-parameters and topology from the operation sequence . With these techniques, NASPY is able to extract the complete NAS model architecture with high fidelity and automation, which are rarely analyzed before.

****

Fact-driven Logical Reasoning

Siru Ouyang,Zhuosheng Zhang,hai zhao

Recent years have witnessed an increasing interest in training machines with reasoning ability, which deeply relies on accurate, clearly presented clue forms that are usually modeled as entity-like knowledge in existing studies. However, in real hierarchical reasoning motivated machine reading comprehension, such one-sided modeling is insufficient for those indispensable local complete facts or events when only "global" knowledge is really paid attention to. Thus, in view of language being a complete knowledge/clue carrier, we propose a general formalism to support representing logic units by extracting backbone constituents of the sentence such as the subject-verb-object formed "facts", covering both global and local knowledge pieces that are necessary as the basis for logical reasoning. Beyond building the ad-hoc graphs, we propose a more general and convenient fact-driven approach to construct a supergraph on top of our newly defined fact units, benefiting from both sides of the connections between facts and internal knowledge such as concepts or actions inside a fact. Experiments on two challenging logical reasoning benchmarks show that our proposed model, \textsc{Focal Reasoner}, outperforms the baseline models dramatically and achieves state-of-the-art results.

****

The Infinite Contextual Graph Markov Model

Daniele Castellana,Federico Errica,Davide Bacciu,Alessio Micheli

The Contextual Graph Markov Model is a deep, unsupervised, and probabilistic model for graphs that is trained incrementally on a layer-by-layer basis. As with most Deep Graph Networks, an inherent limitation is the lack of an automatic mechanism to choose the size of each layer's latent representation. In this paper, we circumvent the problem by extending the Contextual Graph Markov Model with Hierarchical Dirichlet Processes. The resulting model for graphs can automatically adjust the complexity of each layer without the need to perform an extensive model selection. To improve the scalability of the method, we introduce a novel approximated inference procedure that better deals with larger graph topologies. The quality of the learned unsupervised representations is then evaluated across a set of eight graph classification tasks, showing competitive performances against end-to-end supervised methods. The analysis is complemented by studies on the importance of depth, hyper-parameters, and compression of the graph embeddings. We believe this to be an important step towards the theoretically grounded and automatic construction of deep probabilistic architectures for graphs.

****

Uncertainty-based out-of-distribution detection requires suitable function space priors

Francesco D'Angelo,Christian Henning

The need to avoid confident predictions on unfamiliar data has sparked interest in out-of-distribution (OOD) detection. It is widely assumed that Bayesian neural networks (BNNs) are well suited for this task, as the endowed epistemic uncertainty should lead to disagreement in predictions on outliers. In this paper, we question this assumption and show that proper Bayesian inference with function space priors induced by neural networks does not necessarily lead to good OOD det ection. To circumvent the use of approximate inference, we start by studying the infinite-width case, where Bayesian inference can be exact due to the correspondence with Gaussian processes. Strikingly, the kernels induced under common architectural choices lead to uncertainties that do not reflect the underlying data generating process and are therefore unsuited for OOD detection. Importantly, we find this OOD behavior to be consistent with the corresponding finite-width networks. Desirable function space properties can be encoded in the prior in weight space, however, this currently only applies to a specified subset of the domain and thus does not inherently extend to OOD data. Finally, we argue that a trade-off between generalization and OOD capabilities might render the application of BNNs for OOD detection undesirable in practice. Overall, our study discloses fundamental problems when naively using BNNs for OOD detection and opens interesting avenues for future research.

****

COptiDICE: Offline Constrained Reinforcement Learning via Stationary Distribution Correction Estimation
Jongmin Lee,Cosmin Paduraru,Daniel J Mankowitz,Nicolas Heess,Doina Precup,Kee-Eung Kim,Arthur Guez
We consider the offline constrained reinforcement learning (RL) problem, in which the agent aims to compute a policy that maximizes expected return while satisfying given cost constraints, learning only from a pre-collected dataset. This problem setting is appealing in many real-world scenarios, where direct interaction with the environment is costly or risky, and where the resulting policy should comply with safety constraints. However, it is challenging to compute a policy that guarantees satisfying the cost constraints in the offline RL setting, since the off-policy evaluation inherently has an estimation error. In this paper, we present an offline constrained RL algorithm that optimizes the policy in the space of the stationary distribution. Our algorithm, COptiDICE, directly estimates the stationary distribution corrections of the optimal policy with respect to returns, while constraining the cost upper bound, with the goal of yielding a cost-conservative policy for actual constraint satisfaction. Experimental results show that COptiDICE attains better policies in terms of constraint satisfaction and return-maximization, outperforming baseline algorithms.

****

SplitRegex: Faster Regex Synthesis via Neural Example Splitting
Su-Hyeon Kim,Hyunjoon Cheon,Yo-Sub Han,Sang-Ki Ko
Due to the practical importance of regular expressions (regexes, for short), there has been a lot of research to automatically generate regexes from positive and negative string examples. A basic idea of learning a regex is a search-and-repair; search for a correct regex and repair it if incorrect. The problem is known to be PSPACE-complete and the main issue is to obtain a regex quickly within a time limit.
While classical regex learning methods do not perform well, recent approaches using deep neural networks show better performance
with respect to the accuracy of the resulting regexes. However, all these approaches including SOTA models are often extremely slow because of the slow searching mechanism, and do not produce desired regexes within a given time limit.
We tackle the problem of learning regexes faster from positive and negative strings by relying on a novel approach called `neural example splitting'. Our approach essentially split up example strings into multiple parts using a neural network trained to group similar substrings from positive strings. This helps to learn a regex faster and, thus, more accurately since we now learn from several short-length strings.
We propose an effective regex synthesis framework called `SplitRegex' that synthesizes subregexes from `split' positive substrings and produces the final regex by concatenating the synthesized subregexes. For the negative sample, we exploit pre-generated subregexes during the subregex synthesis process and perform the matching against negative strings. Then the final regex becomes consistent with all negative strings. SplitRegex is a divided-and-conquer framework for learning target regexes; split (=divide) positive strings and infer partial regexes for

multiple parts, which is much more accurate than the whole string inferring, and concatenate (=conquer) inferred regexes while satisfying negative strings. We empirically demonstrate that the proposed SplitRegex framework substantially improves the previous regex synthesis approaches over four benchmark datasets.

****

Wavelet Feature Maps Compression for Low Bandwidth Convolutional Neural Networks

Yair Zohav,Shahaf E Finder,Maor Ashkenazi,Eran Treister

Quantization is one of the most effective techniques for compressing Convolutional Neural Networks (CNNs), which are known for requiring extensive computational resources. However, aggressive quantization may cause severe degradation in the prediction accuracy of such networks, especially in image-to-image tasks such as semantic segmentation and depth prediction. In this paper, we propose Wavelet Compressed Convolution (WCC)---a novel approach for activation maps compression for 1×1 convolutions (the workhorse of modern CNNs). WCC achieves compression ratios and computational savings that are equivalent to low bit quantization rates at a relatively minimal loss of accuracy. To this end, we use a hardware-friendly Haar-wavelet transform, known for its effectiveness in image compression, and define the convolution on the compressed activation map. WCC can be utilized with any 1×1 convolution in an existing network architecture. By combining WCC with light quantization, we show that we achieve compression rates equal to 2-bit and 1-bit with minimal degradation in image-to-image tasks.

****

Learning Dynamics Models for Model Predictive Agents

Michael Lutter,Leonard Hasenclever,Arunkumar Byravan,Gabriel Dulac-Arnold,Piotr Trochim,Nicolas Heess,Josh Merel,Yuval Tassa

Model-Based Reinforcement Learning involves learning a dynamics model from data, and then using this model to optimise behaviour, most often with an online planner. Much of the recent research along these lines presents a particular set of design choices, involving problem definition, model learning and planning. Given the multiple contributions, it is difficult to evaluate the effects of each. This paper sets out to disambiguate the role of different design choices for learning dynamics models, by comparing their performance to planning with a ground-truth model -- the simulator. First, we collect a rich dataset from the training sequence of a model-free agent on 5 domains of the DeepMind Control Suite. Second, we train feed-forward dynamics models in a supervised fashion, and evaluate planner performance while varying and analysing different model design choices, including ensembling, stochasticity, multi-step training and timestep size. Besides the quantitative analysis, we describe a set of qualitative findings, rules of thumb, and future research directions for planning with learned dynamics models. Videos of the results are available at https://sites.google.com/view/learning-better-models.

****

Pretrained models are active learners

Alex Tamkin,Dat Nguyen,Salil Deshpande,Jesse Mu,Noah Goodman

An important barrier to the safe deployment of machine learning systems is the risk of \emph{task ambiguity}, where multiple behaviors are consistent with the provided examples. We investigate whether pretrained models are better active learners, capable of asking for example labels that \textit{disambiguate} between the possible tasks a user may be trying to specify. Across a range of image and text datasets with spurious correlations, latent minority groups, or domain shifts, finetuning pretrained models with data acquired through simple uncertainty sampling achieves the same accuracy with \textbf{up to 6$\times$ fewer labels} compared to random sampling. Moreover, the examples chosen by these models are preferentially minority classes or informative examples where the spurious feature and class label are decorrelated. Notably, gains from active learning are not seen in unpretrained models, which do not select such examples, suggesting that the ability to actively learn is an emergent property of the pretraining process.

****

On Margin Maximization in Linear and ReLU Networks

Gal Vardi,Ohad Shamir,Nathan Srebro

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 the 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. Finally, we answer a question posed in Lyu and Li [2019] by showing that for non-homogeneous networks, the normalized margin may strictly decrease over time.

****

Gradient-based Meta-solving and Its Applications to Iterative Methods for Solving Differential Equations

Sohei Arisaka,Qianxiao Li

In science and engineering applications, it is often required to solve similar computational problems repeatedly. In such cases, we can utilize the data from previously solved problem instances to improve efficiency of finding subsequent solutions. This offers a unique opportunity to combine machine learning (in particular, meta-learning) and scientific computing. To date, a variety of such domain-specific methods have been proposed in the literature, but a generic approach for designing these methods remains under-explored. In this paper, we tackle this issue by formulating a general framework to describe these problems, and propose a gradient-based algorithm to solve them in a unified way. As an illustration of this approach, we study the adaptive generation of initial guesses for iterative solvers to speed up the solution of differential equations. We demonstrate the performance and versatility of our method through theoretical analysis and numerical experiments.

****

GRODIN: Improved Large-Scale Out-of-Domain detection via Back-propagation

Gleb Yengalych,Igor E. Kuralenok,Vasily A Ershov

Uncertainty estimation and out-of-doman (OOD) input detection are critical for improving the safety and robustness of machine learning. Unfortunately, most methods for detecting OOD examples have been evaluated on small tasks while typical methods are computationally expensive. In this paper we propose a new gradient-based method called GRODIN for OOD detection. The proposed method is conceptually simple, computationally cheaper than ensemble methods and can be directly applied to any existing and deployed model without re-training. We evaluate GRODIN on models trained on CIFAR-10 and ImageNet datasets, and show it's strong performance on various OOD ImageNet datasets such as ImageNet-O, ImageNet-A, ImageNet-R, ImageNet-C.

****

AIR-Net: Adaptive and Implicit Regularization Neural Network for matrix completion

Zhemin Li,Hongxia Wang

Conventionally, the matrix completion (MC) model aims to recover a matrix from partially observed elements. Accurate recovery necessarily requires a regularization encoding priors of the unknown matrix/signal properly. However, encoding the priors accurately for the complex natural signal is difficult, and even then, the model might not generalize well outside the particular matrix type. This work combines adaptive and implicit low-rank regularization that captures the prior dynamically according to the current recovered matrix. Furthermore, we aim to answer the question: how does adaptive regularization affect implicit regularization? We utilize neural networks to represent Adaptive and Implicit Regularization and named the proposed model \textit{AIR-Net}. Theoretical analyses show that the adaptive part of the AIR-Net enhances implicit regularization. In addition, the adaptive regularizer vanishes at the end, thus can avoid saturation issues. Numerical experiments for various data demonstrate the effectiveness of AIR-Net, especially when the locations of missing elements are not randomly chosen. With complete flexibility to select neural networks for matrix representation, AIR-Net can be extended to solve more general inverse problems.

****

Equivariant Vector Field Network for Many-body System Modeling

weitao Du,He Zhang,Yuanqi Du,Qi Meng,Wei Chen,Bin Shao,Tie-Yan Liu

Modeling many-body systems has been a long-standing challenge in science, from classical and quantum physics to computational biology. Equivariance is a critical physical symmetry for many-body dynamic systems, which enables robust and accurate prediction under arbitrary reference transformations. In light of this, great efforts have been put on encoding this symmetry into deep neural networks, which significantly boosts the prediction performance of down-streaming tasks. Some general equivariant models which are computationally efficient have been proposed, however, these models have no guarantee on the approximation power and may have information loss. In this paper, we leverage insights from the scalarization technique in differential geometry to model many-body systems by learning the gradient vector fields, which are SE(3) and permutation equivariant. Specifically, we propose the Equivariant Vector Field Network (EVFN), which is built on a novel tuple of equivariant basis and the associated scalarization and vectorization layers. Since our tuple equivariant basis forms a complete basis, learning the dynamics with our EVFN has no information loss. We evaluate our method on predicting trajectories of simulated Newton mechanics systems with both full and partially observed data, as well as the equilibrium state of small molecules (molecular conformation) evolving as a statistical mechanics system. Experimental results across multiple tasks demonstrate that our model achieves best or competitive performance on baseline models in various types of datasets.

****

Objects in Semantic Topology

Shuo Yang,Peize Sun,Yi Jiang,Xiaobo Xia,Ruiheng Zhang,Zehuan Yuan,Changhu Wang,Ping Luo,Min Xu

A more realistic object detection paradigm, Open-World Object Detection, has arised increasing research interests in the community recently. A qualified open-world object detector can not only identify objects of known categories, but also discover unknown objects, and incrementally learn to categorize them when their annotations progressively arrive. Previous works rely on independent modules to recognize unknown categories and perform incremental learning, respectively. In this paper, we provide a unified perspective: Semantic Topology. During the life-long learning of an open-world object detector, all object instances from the same category are assigned to their corresponding pre-defined node in the semantic topology, including the `unknown' category. This constraint builds up discriminative feature representations and consistent relationships among objects, thus enabling the detector to distinguish unknown objects out of the known categories, as well as making learned features of known objects undistorted when learning new categories incrementally. Extensive experiments demonstrate that semantic topology, either randomly-generated or derived from a well-trained language model, could outperform the current state-of-the-art open-world object detectors by a large margin, e.g., the absolute open-set error (the number of unknown instances that are wrongly labeled as known) is reduced from 7832 to 2546, exhibiting the inherent superiority of semantic topology on open-world object detection.

****

Adaptive Generalization for Semantic Segmentation

Sherwin Bahmani,Oliver Hahn,Eduard Sebastian Zamfir,Nikita Araslanov,Stefan Roth

Out-of-distribution robustness remains a salient weakness of current state-of-the-art models for semantic segmentation. Until recently, research on generalization followed a restrictive assumption that the model parameters remain fixed after the training process. In this work, we empirically study an adaptive inference strategy for semantic segmentation that adjusts the model to the test sample before producing the final prediction. We achieve this with two complementary techniques. Using Instance-adaptive Batch Normalization (IaBN), we modify normalization layers by combining the feature statistics acquired at training time with those of the test sample. We next introduce a test-time training (TTT) approach fo r semantic segmentation, Seg-TTT, which adapts the model parameters to the test sample using a self-supervised loss. Relying on a more rigorous evaluation protocol compared to previous work on generalization in semantic segmentation, our study shows that these techniques consistently and significantly outperform the baseline and attain a new state of the art, substantially improving in accuracy over previous generalization methods.

****

Context-invariant, multi-variate time series representations

Stephan Rabanser,Tim Januschowski,Kashif Rasul,Oliver Borchert,Richard Kurle,Jan Gasthaus,Michael Bohlke-Schneider,Nicolas Papernot,Valentin Flunkert

Modern time series corpora, in particular those coming from sensor-based data, exhibit characteristics that have so far not been adequately addressed in the literature on representation learning for time series. In particular, such corpora often allow to distinguish between \emph{exogenous} signals that describe a context which influences a given appliance and \emph{endogenous} signals that describe the internal state of the appliance. We propose a temporal convolution network based embedding that improves on the state-of-the-art by incorporating recent advances in contrastive learning to the time series domain and by adopting a multi-resolution approach. Employing techniques borrowed from domain-adversarial learning, we achieve an invariance of the embeddings with respect to the context provided by the exogenous signal. To show the effectiveness of our approach, we contribute new data sets to the research community and use both new as well as existing data sets to empirically verify that we can separate normal from abnormal internal appliance behaviour independent of the external signals in data sets from IoT and DevOps.

****

Hidden Parameter Recurrent State Space Models For Changing Dynamics Scenarios

Vaisakh Shaj,Dieter Büchler,Rohit Sonker,Philipp Becker,Gerhard Neumann

Recurrent State-space models (RSSMs) are highly expressive models for learning patterns in time series data and for system identification. However, these models are often based on the assumption that the dynamics are fixed and unchanging, which is rarely the case in real-world scenarios. Many control applications often exhibit tasks with similar, but not identical dynamics, that can be modelled as having a common latent structure. We introduce the Hidden Parameter Recurrent State Space Models (HiP-RSSMs), a framework that parametrizes a family of related state-space models with a low-dimensional set of latent factors. We present a simple and effective way of performing learning and inference over this Gaussian graphical model that avoids approximations like variational inference. We show that HiP-RSSMs outperforms RSSMs and competing multi-task models on several challenging robotic benchmarks both on real systems and simulations.

****

Graph Neural Network Guided Local Search for the Traveling Salesperson Problem

Benjamin Hudson,Qingbiao Li,Matthew Malencia,Amanda Prorok

Solutions to the Traveling Salesperson Problem (TSP) have practical applications to processes in transportation, logistics, and automation, yet must be computed with minimal delay to satisfy the real-time nature of the underlying tasks. However, solving large TSP instances quickly without sacrificing solution quality remains challenging for current approximate algorithms. To close this gap, we present a hybrid data-driven approach for solving the TSP based on Graph Neural Networks (GNNs) and Guided Local Search (GLS). Our model predicts the regret of including each edge of the problem graph in the solution; GLS uses these predictions in conjunction with the original problem graph to find solutions. Our experiments demonstrate that this approach converges to optimal solutions at a faster rate than three recent learning based approaches for the TSP. Notably, we reduce the mean optimality gap on the 100-node problem set from 1.534% to 0.705%, a 2x improvement. When generalizing from 20-node instances to the 100-node problem set, we reduce the optimality gap from 18.845% to 2.622%, a 7x improvement.

****

DiBB: Distributing Black-Box Optimization

Giuseppe Cuccu,Luca Sven Rolshoven,Fabien Vorpe,Philippe Cudre-Mauroux,Tobias Gl asmachers

We present a novel framework for Distributing Black-Box Optimization (DiBB). DiBB can encapsulate any Black Box Optimization (BBO) method, making it of particular interest for scaling and distributing modern Evolution Strategies (ES), such as CMA-ES and its variants, which maintain a sampling covariance matrix throughout the run. Due to high algorithmic complexity however, such methods are unsuitable alone to address high-dimensional problems, e.g. for sophisticated Reinforcement Learning (RL) control. This limits the applicable methods to simpler ES, which trade off faster updates for lowered sample efficiency. DiBB overcomes this limitation by means of problem decomposition, leveraging expert knowledge in the problem structure such as a known topology for a neural network controller. This allows to distribute the workload across an arbitrary number of nodes in a cluster, while maintaining the feasibility of second order (covariance) learning on high-dimensional problems. The computational complexity per node is bounded by the (arbitrary) size of blocks of variables, which is independent of the problem size.

****

Subspace State-Space Identification and Model Predictive Control of Nonlinear Dynamical Systems Using Deep Neural Network with Bottleneck

Ichiro Maruta,Keito Yamada,Kenji Fujimoto

A novel nonlinear system identification method that produces state estimator and predictor directly usable for model predictive control (MPC) is proposed in this paper. The main feature of the proposed method is that it uses a neural network with a bottleneck layer between the state estimator and predictor to represent the input-output dynamics, and it is proven that the state of the dynamical system can be extracted from the bottleneck layer based on the observability of the target system. The training of the network is shown to be a natural nonlinear extension of the subspace state-space system identification method established for linear dynamical systems. This correspondence gives interpretability to the resulting model based on linear control theory. The usefulness of the proposed method and the interpretability of the model are demonstrated through an illustrative example of MPC.

****

One Thing to Fool them All: Generating Interpretable, Universal, and Physically-Realizable Adversarial Features

Stephen Casper,Max Nadeau,Gabriel Kreiman

It is well understood that modern deep networks are vulnerable to adversarial attacks. However, conventional methods fail to produce adversarial perturbations that are intelligible to humans, and they pose limited threats in the physical world. To study feature-class associations in networks and better understand the real-world threats they face, we develop feature-level adversarial perturbations using deep image generators and a novel optimization objective. We term these feature-fool attacks. We show that they are versatile and use them to generate targeted feature-level attacks at the ImageNet scale that are simultaneously interpretable, universal to any source image, and physically-realizable. These attacks can also reveal spurious, semantically-describable feature/class associations, and we use them to guide the design of ``copy/paste'' adversaries in which one natural image is pasted into another to cause a targeted misclassification.

****

Deep Active Learning by Leveraging Training Dynamics

Haonan Wang,Wei Huang,Hanghang Tong,Andrew J Margenot,Jingrui He

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 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. Further on, 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.
****

On the Pitfalls of Heteroscedastic Uncertainty Estimation with Probabilistic Neural Networks

Maximilian Seitzer,Arash Tavakoli,Dimitrije Antic,Georg Martius

Capturing aleatoric uncertainty is a critical part of many machine learning systems. In deep learning, a common approach to this end is to train a neural network to estimate the parameters of a heteroscedastic Gaussian distribution by maximizing the logarithm of the likelihood function under the observed data. In this work, we examine this approach and identify potential hazards associated with the use of log-likelihood in conjunction with gradient-based optimizers. First, we present a synthetic example illustrating how this approach can lead to very poor but stable parameter estimates. Second, we identify the culprit to be the log-likelihood loss, along with certain conditions that exacerbate the issue. Third, we present an alternative formulation, termed β-NLL, in which each data point's contribution to the loss is weighted by the β-exponentiated variance estimate. We show that using an appropriate β largely mitigates the issue in our illustrative example. Fourth, we evaluate this approach on a range of domains and tasks and show that it achieves considerable improvements and performs more robustly concerning hyperparameters, both in predictive RMSE and log-likelihood criteria.
****

Modelling neuronal behaviour with time series regression: Recurrent Neural Netwoks on synthetic C. elegans data

Gonçalo Leote Cardoso Mestre,Ruxandra Barbulescu,Arlindo L. Oliveira,L. Miguel Silveira

Given the inner complexity of the human nervous system, insight into the dynamics of brain activity can be gained from understanding smaller and simpler organisms, such as the nematode C. elegans. The behavioural and structural biology of these organisms is well-known, making them prime candidates for benchmarking modelling and simulation techniques. In these complex neuronal collections, classical white-box modelling techniques based on intrinsic structural or behavioural information are either unable to capture the profound nonlinearities of the neuronal response to different stimuli or generate extremely complex models, which are computationally intractable. In this paper we investigate whether it is possible to generate lower complexity black-box models that can capture the system dynamics with low error using only measured or simulated input-output information. We show how the nervous system of C. elegans can be modelled and simulated with data-driven models using different neural network architectures. Specifically, we target the use of state of the art recurrent neural networks architectures such as LSTMs and GRUs and compare these architectures in terms of their properties and their RMSE, as well as the complexity of the resulting models.
We show that GRU models with a hidden layer size of 4 units are able to accurately reproduce the system's response to very different stimuli.
****

Knothe-Rosenblatt transport for Unsupervised Domain Adaptation

Aladin Virmaux,Illyyne Saffar,Jianfeng Zhang,Balázs Kégl

Unsupervised domain adaptation (UDA) aims at exploiting related but different data sources in order to tackle a common task in a target domain. UDA remains a central yet challenging problem in machine learning.

In this paper, we present an approach based on the Knothe-Rosenblatt transport: we exploit autoregressive density estimation algorithms to accurately model the different sources by an autoregressive model using a mixture of Gaussians.

Our Knothe-Rosenblatt Domain Adaptation (KRDA) then takes advantage of the triangularity of the autoregressive models to build an explicit mapping of the source samples into the target domain. We show that the transfer map built by KRDA preserves each component quantiles of the observations, hence aligning the representations of the different data sets in the same target domain.

Finally, we show that KRDA has state-of-the-art performance on both synthetic and real world UDA problems.

****

Detecting Modularity in Deep Neural Networks

Shlomi Hod,Stephen Casper,Daniel Filan,Cody Wild,Andrew Critch,Stuart Russell

A neural network is modular to the extent that parts of its computational graph (i.e. structure) can be represented as performing some comprehensible subtask relevant to the overall task (i.e. functionality). Are modern deep neural networks modular? How can this be quantified? In this paper, we consider the problem of assessing the modularity exhibited by a partitioning of a network's neurons. We propose two proxies for this: importance, which reflects how crucial sets of neurons are to network performance; and coherence, which reflects how consistently their neurons associate with features of the inputs. To measure these proxies, we develop a set of statistical methods based on techniques conventionally used to interpret individual neurons. We apply the proxies to partitionings generated by spectrally clustering a graph representation of the network's neurons with edges determined either by network weights or correlations of activations. We show that these partitionings, even ones based only on weights (i.e. strictly from non-runtime analysis), reveal groups of neurons that are important and coherent. These results suggest that graph-based partitioning can reveal modularity and help us understand how deep neural networks function.

****

Feature Flow Regularization: Improving Structured Sparsity in Deep Neural Networks

Yue Wu,Yuan Lan,Luchan Zhang,Yang Xiang

Pruning is a model compression method that removes redundant parameters and accelerates the inference speed of deep neural networks (DNNs) while maintaining accuracy. Most available pruning methods impose various conditions on parameters or features directly. In this paper, we propose a simple and effective regularization strategy to improve the structured sparsity and structured pruning in DNNs from a new perspective of evolution of features. In particular, we consider the trajectories connecting features of adjacent hidden layers, namely feature flow. We propose feature flow regularization (FFR) to penalize the length and the total absolute curvature of the trajectories, which implicitly increases the structured sparsity of the parameters. The principle behind FFR is that short and straight trajectories will lead to an efficient network that avoids redundant parameters. Experiments on CIFAR-10 and ImageNet datasets show that FFR improves structured sparsity and achieves pruning results comparable to or even better than those state-of-the-art methods.

****

Label-Efficient Semantic Segmentation with Diffusion Models

Dmitry Baranchuk,Andrey Voynov,Ivan Rubachev,Valentin Khrulkov,Artem Babenko

Denoising diffusion probabilistic models have recently received much research attention since they outperform alternative approaches, such as GANs, and currently provide state-of-the-art generative performance. The superior performance of diffusion models has made them an appealing tool in several applications, including inpainting, super-resolution, and semantic editing. In this paper, we demonstrate that diffusion models can also serve as an instrument for semantic segmentation, especially in the setup when labeled data is scarce. In particular, for several pretrained diffusion models, we investigate the intermediate activations from the networks that perform the Markov step of the reverse diffusion process. We show that these activations effectively capture the semantic information from an input image and appear to be excellent pixel-level representations for the segmentation problem. Based on these observations, we describe a simple segmentation method, which can work even if only a few training images are provided. Our approach significantly outperforms the existing alternatives on several datasets for the same amount of human supervision.

****

Towards fast and effective single-step adversarial training

Pau de Jorge,Adel Bibi,Riccardo Volpi,Amartya Sanyal,Philip Torr,Grégory Rogez,Puneet K. Dokania

Recently, Wong et al. (2020) showed 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. Moreover, they showed adding a random perturbation prior to FGSM (RS-FGSM) seemed to be sufficient to prevent CO. However, Andriushchenko & Flammarion (2020) observed that RS-FGSM still leads to CO for larger perturbations and argue that the only contribution of the random step is to reduce the magnitude of the attacks. They suggest a regularizer (GradAlign) that avoids CO but is significantly more expensive than RS-FGSM. In this work, we methodically revisit the role of noise and clipping in single-step adversarial training. Contrary to previous intuitions, we find that not clipping the perturbation around the clean sample and using a stronger noise is highly effective in avoiding CO for large perturbation radii, despite leading to an increase in the magnitude of the attacks. Based on these observations, we propose a method called Noise-FGSM (N-FGSM), which attacks noise-augmented samples directly using a single-step. Empirical analyses on a large suite of experiments show that N-FGSM is able to match or surpass the performance of GradAlign while achieving a 3x speed-up.

****

Long Document Summarization with Top-Down and Bottom-Up Representation Inference

Bo Pang,Erik Nijkamp,Wojciech Maciej Kryscinski,Silvio Savarese,Yingbo Zhou,Caiming Xiong

Text summarization aims to condense long documents and retain key information. Critical to the success of a summarization model is the faithful inference of latent representations of words or tokens in the source documents. Most recent models infer the latent representations with a transformer encoder, which is purely bottom-up. Also, self-attention-based inference models face the challenge of quadratic complexity with respect to sequence length. We propose a principled inference framework to improve summarization models on these two aspects. Our framework assumes a hierarchical latent structure of a document where the top-level captures the long range dependency at a coarser time scale and the bottom token level preserves the details. Critically, this hierarchical structure enables token representations to be updated in both a bottom-up and top-down manner. In the bottom-up pass, token representations are inferred with local self-attention to leverage its efficiency. Top-down correction is then applied to allow tokens to capture long-range dependency. We demonstrate the effectiveness of the proposed framework on a diverse set of summarization datasets, including narrative, conversational, scientific documents and news. Our model achieves (1) competitive or better performance on short documents with higher memory and compute efficiency, compared to full attention transformers, and (2) state-of--the-art performance on a wide range of long document summarization benchmarks, compared to recent efficient transformers. We also show that our model can summarize an entire book and achieve competitive performance using 0.27% parameters (464M vs. 175B) and much less training data, compared to a recent GPT-3-based model. These results indicate the general applicability and benefits of the proposed framework.

****

Function-Space Variational Inference for Deep Bayesian Classification

Jihao Andreas Lin,Joe Watson,Pascal Klink,Jan Peters

Bayesian deep learning approaches assume model parameters to be latent random variables and infer posterior predictive distributions to quantify uncertainty, increase safety and trust, and prevent overconfident and unpredictable behavior. However, weight-space priors are model-specific, can be difficult to interpret and hard to choose. Instead of weight-space priors, we leverage function-space variational inference to apply a Dirichlet predictive prior in function space, resulting in a variational Dirichlet posterior which facilitates easier specificatio n of epistemic uncertainty. This is achieved through the perspective of stochastic neural network classifiers as variational implicit processes, which can be trained using function-space variational inference by devising a novel Dirichlet KL estimator. Experiments on small- and large-scale image classification tasks demonstrate that our function-space inference scales to large-scale tasks and models, improves adversarial robustness and boosts uncertainty quantification across models, without influencing the in-distribution performances, architecture or model size.

****

SALT : Sharing Attention between Linear layer and Transformer for tabular dataset

Juseong Kim,Jinsun Park,Giltae Song

Handling tabular data with deep learning models is a challenging problem despite their remarkable success in vision and language processing applications. Therefore, many practitioners still rely on classical models such as gradient boosting decision trees (GBDTs) rather than deep networks due to their superior performance with tabular data. In this paper, we propose a novel hybrid deep network architecture for tabular data, dubbed SALT (Sharing Attention between Linear layer and Transformer). The proposed SALT consists of two blocks: Transformers and linear layers blocks that take advantage of shared attention matrices. The shared attention matrices enable transformers and linear layers to closely cooperate with each other, and it leads to improved performance and robustness. Our algorithm outperforms tree-based ensemble models and previous deep learning methods in multiple benchmark datasets. We further demonstrate the robustness of the proposed SALT with semi-supervised learning and pre-training with small dataset scenarios.

****

Scalable Hierarchical Embeddings of Complex Networks

Nikolaos Nakis,Abdulkadir CELIKKANAT,Sune Lehmann,Morten Mørup

Graph representation learning has become important in order to understand and predict intrinsic structures in complex networks. A variety of embedding methods has in recent years been developed including the Latent Distance Modeling (LDM) approach. A major challenge is scaling network embedding approaches to very large networks and a drawback of LDM is the computational cost invoked evaluating the full likelihood having O(N^2) complexity, making such analysis of large networks infeasible. We propose a novel multiscale hierarchical estimate of the full likelihood of LDMs providing high details where the likelihood approximation is most important while scaling in complexity at O(NlogN). The approach relies on a clustering procedure approximating the Euclidean norm of every node pair according to the multiscale hierarchical structure imposed. We demonstrate the accuracy of our approximation and for the first time embed very large networks in the order of a million nodes using LDM and contrast the predictive performance to prominent scalable graph embedding approaches. We find that our approach significantly outperforms these existing scalable approaches in the ability to perform link prediction, node clustering, and classification utilizing a surprisingly low embedding dimensionality of two to three dimensions whereas the extracted hierarchical structure facilitates network visualization and interpretation. The developed scalable hierarchical embedding approach enables accurate low dimensional representations of very large networks providing detailed visualizations that can further our understanding of their properties and structure.

****

CoLLIE: Continual Learning of Language Grounding from Language-Image Embeddings

Gabriel Skantze,Bram Willemsen

This paper presents CoLLIE: a simple, yet effective model for continual learning of how language is grounded in vision. Given a pre-trained multimodal embedding model, where language and images are projected in the same semantic space (in this case CLIP by OpenAI), CoLLIE learns a transformation function that adjusts the language embeddings when needed to accommodate new language use. Unlike traditional few-shot learning, the model does not just learn new classes and labels, but can also generalize to similar language use. We verify the model's performan

ce on two different tasks of continual learning and show that it can efficiently learn and generalize from only a few examples, with little interference with the model's original zero-shot performance.
****

Two Sides of the Same Coin: Heterophily and Oversmoothing in Graph Convolutional Neural Networks
Yujun Yan,Milad Hashemi,Kevin Swersky,Yaoqing Yang,Danai Koutra
In node classification tasks, heterophily and oversmoothing are two problems that can hurt the performance of graph convolutional neural networks (GCNs). The heterophily problem refers to the model's inability to handle heterophilous graphs where neighboring nodes belong to different classes; the oversmoothing problem refers to the model's degenerated performance with increasing number of layers. These two seemingly unrelated problems have been studied mostly independently, but there is recent empirical evidence that solving one problem may benefit the other.

In this work, beyond empirical observations, we aim to: (1) analyze the heterophily and oversmoothing problems from a unified theoretical perspective, (2) identify the common causes of the two problems based on our theories, and (3) propose simple yet effective strategies to address the common causes. In our theoretical analysis, we show that the common causes of the heterophily and oversmoothing problems---namely, the relative degree of a node (compared to its neighbors) and its heterophily level---trigger the node representations in consecutive layers to "move" closer to the original decision boundary, which increases the misclassification rate of node labels under certain constraints. We theoretically show that: (1) Nodes with high heterophily have a higher misclassification rate. (2) Even with low heterophily, degree disparity in a node's neighborhood can influence the movements of node representations and result in a "pseudo-heterophily" situation, which helps to explain oversmoothing. (3) Allowing not only positive, but also negative messages during message passing can help counteract the common causes of the two problems. Based on our theoretical insights, we propose simple modifications to the GCN architecture (i.e., learned degree corrections and signed messages), and we show that they alleviate the heteorophily and oversmoothing problems with extensive experiments on nine real networks. Compared to other approaches, which tend to work well in either heterophily or oversmoothing, our modified GCN model performs well in both problems.
****

Language model compression with weighted low-rank factorization
Yen-Chang Hsu,Ting Hua,Sungen Chang,Qian Lou,Yilin Shen,Hongxia Jin
Factorizing a large matrix into small matrices is a popular strategy for model compression. Singular value decomposition (SVD) plays a vital role in this compression strategy, approximating a learned matrix with fewer parameters. However, SVD minimizes the squared error toward reconstructing the original matrix without gauging the importance of the parameters, potentially giving a larger reconstruction error for those who affect the task accuracy more. In other words, the optimization objective of SVD is not aligned with the trained model's task accuracy. We analyze this previously unexplored problem, make observations, and address it by introducing Fisher information to weigh the importance of parameters affecting the model prediction. This idea leads to our method: Fisher-Weighted SVD (FWSVD). Although the factorized matrices from our approach do not result in smaller reconstruction errors, we find that our resulting task accuracy is much closer to the original model's performance. We perform analysis with the transformer-based language models, showing our weighted SVD largely alleviates the mismatched optimization objectives and can maintain model performance with a higher compression rate. Our method can directly compress a task-specific model while achieving better performance than other compact model strategies requiring expensive model pre-training. Moreover, the evaluation of compressing an already compact model shows our method can further reduce 9% to 30% parameters with an insignificant impact on task accuracy.
****

Learning Vision-Guided Quadrupedal Locomotion End-to-End with Cross-Modal Transformers

Ruihan Yang,Minghao Zhang,Nicklas Hansen,Huazhe Xu,Xiaolong Wang

We propose to address quadrupedal locomotion tasks using Reinforcement Learning (RL) with a Transformer-based model that learns to combine proprioceptive information and high-dimensional depth sensor inputs. While learning-based locomotion has made great advances using RL, most methods still rely on domain randomization for training blind agents that generalize to challenging terrains. Our key insight is that proprioceptive states only offer contact measurements for immediate reaction, whereas an agent equipped with visual sensory observations can learn to proactively maneuver environments with obstacles and uneven terrain by anticipating changes in the environment many steps ahead. In this paper, we introduce LocoTransformer, an end-to-end RL method that leverages both proprioceptive states and visual observations for locomotion control. We evaluate our method in challenging simulated environments with different obstacles and uneven terrain. We transfer our learned policy from simulation to a real robot by running it indoor and in-the-wild with unseen obstacles and terrain. Our method not only significantly improves over baselines, but also achieves far better generalization performance, especially when transferred to the real robot. Our project page with videos is at https://rchalyang.github.io/LocoTransformer/.

****

Neural Implicit Representations for Physical Parameter Inference from a Single Video

Florian Hofherr,Lukas Koestler,Florian Bernard,Daniel Cremers

Neural networks have recently been used to model the dynamics of diverse physical systems. While existing methods achieve impressive results, they are limited by their strong demand for training data and their weak generalization abilities. To overcome these limitations, in this work we propose to combine neural implicit representations for appearance modeling with neural ordinary differential equations (ODEs) in order to obtain interpretable physical models directly from visual observations. Our proposed model combines several unique advantages: (i) It is trained from a single video, and thus overcomes the need for large training datasets. (ii) The use of neural implicit representation enables the processing of high-resolution videos and the synthesis of photo-realistic imagery. (iii) The embedded neural ODE has a known parametric form that allows for the identification of interpretable physical parameters, and (iv) long-term prediction in state space. (v) Furthermore, the photo-realistic rendering of novel scenes with modified physical parameters becomes possible.

****

Pareto Set Learning for Neural Multi-Objective Combinatorial Optimization

Xi Lin,Zhiyuan Yang,Qingfu Zhang

Multiobjective combinatorial optimization (MOCO) problems can be found in many real-world applications. However, exactly solving these problems would be very challenging, particularly when they are NP-hard. Many handcrafted heuristic methods have been proposed to tackle different MOCO problems over the past decades. In this work, we generalize the idea of neural combinatorial optimization, and develop a learning-based approach to approximate the whole Pareto set for a given MOCO problem without further search procedure. We propose a single preference-conditioned model to directly generate approximate Pareto solutions for any trade-off preference, and design an efficient multiobjective reinforcement learning algorithm to train this model. Our proposed method can be treated as a learning-based extension for the widely-used decomposition-based multiobjective evolutionary algorithm (MOEA/D). It uses a single model to accommodate all the possible preferences, whereas other methods use a finite number of solutions to approximate the Pareto set. Experimental results show that our proposed method significantly outperforms some other methods on the multiobjective traveling salesman problem, multiobjective vehicle routing problem, and multiobjective knapsack problem in terms of solution quality, speed, and model efficiency.

****

Prototypical Contrastive Predictive Coding

Kyungmin Lee

Transferring representational knowledge of a model to another is a wide-ranging topic in machine learning. Those applications include the distillation of a large supervised or self-supervised teacher model to a smaller student model or self-supervised learning via self-distillation. Knowledge distillation is an original method to solve these problems, which minimizes a cross-entropy loss between the prototypical probabilistic outputs of teacher and student networks. On the other hand, contrastive learning has shown its competency in transferring representations as they allow students to capture the information of teacher representations. In this paper, we amalgamate the advantages of knowledge distillation and contrastive learning by modeling the critic of a contrastive objective by the prototypical probabilistic discrepancy between two features. We refer to it as prototypical contrastive predictive coding and present efficient implementation using the proposed objective for three distillation tasks: supervised model compression, self-supervised model compression, and self-supervised learning via self-distillation. Through extensive experiments, we validate the effectiveness of our method and show that our method achieves state-of-the-art performance in supervised / self-supervised model compression.

****

Adaptive Q-learning for Interaction-Limited Reinforcement Learning

Han Zheng,Xufang Luo,pengfei wei,Xuan Song,Dongsheng Li,Jing Jiang

Conventional reinforcement learning (RL) needs an environment to collect fresh data, which is impractical when an online interaction is costly.

Offline RL provides an alternative solution by directly learning from the logged dataset. However, it usually yields unsatisfactory performance due to a pessimistic update scheme or/and the low quality of logged datasets.

Moreover, how to evaluate the policy under the offline setting is also a challenging problem.

In this paper, we propose a unified framework called Adaptive Q-learning for effectively taking advantage of offline and online learning.

Specifically, we explicitly consider the difference between the online and offline data and apply an adaptive update scheme accordingly, i.e., a pessimistic update strategy for the offline dataset and a greedy or no pessimistic update scheme for the online dataset.

When combining both, we can apply very limited online exploration steps to achieve expert performance even when the offline dataset is poor, e.g., random dataset.

Such a framework provides a unified way to mix the offline and online RL and gain the best of both worlds.

To understand our framework better, we then provide an initialization following our framework's setting.

Extensive experiments are done to verify the effectiveness of our proposed method.

****

Adversarial Robustness Through the Lens of Causality

Yonggang Zhang,Mingming Gong,Tongliang Liu,Gang Niu,Xinmei Tian,Bo Han,Bernhard Schölkopf,Kun Zhang

The adversarial vulnerability of deep neural networks has attracted significant attention in machine learning. As causal reasoning has an instinct for modeling distribution change, it is essential to incorporate causality into analyzing this specific type of distribution change induced by adversarial attacks. However, causal formulations of the intuition of adversarial attacks and the development of robust DNNs are still lacking in the literature. To bridge this gap, we construct a causal graph to model the generation process of adversarial examples and define the adversarial distribution to formalize the intuition of adversarial attacks. From the causal perspective, we study the distinction between the natural and adversarial distribution and conclude that the origin of adversarial vulnerability is the focus of models on spurious correlations. Inspired by the causal understanding, we propose the \emph{Causal}-inspired \emph{Adv}ersarial distribution alignment method, CausalAdv, to eliminate the difference between natural and adversarial distributions by considering spurious correlations. Extensive experiments demonstrate the efficacy of the proposed method. Our work is the first attempt towards using causality to understand and mitigate the adversarial vulnerability.

****

Practical and Private Heterogeneous Federated Learning

Hanxiao Chen,Meng Hao,Hongwei Li,Guangxiao Niu,Guowen Xu,Huawei Wang,Yuan Zhang,Tianwei Zhang

Heterogeneous federated learning (HFL) enables clients with different computation/communication capabilities to collaboratively train their own customized models, in which the knowledge of models is shared via clients' predictions on a public dataset. However, there are two major limitations: 1) The assumption of public datasets may be unrealistic for data-critical scenarios such as Healthcare and Finance. 2) HFL is vulnerable to various privacy violations since the samples and predictions are completely exposed to adversaries. In this work, we develop PrivHFL, a general and practical framework for privacy-preserving HFL. We bypass the limitations of public datasets by designing a simple yet effective dataset expansion method. The main insight is that expanded data could provide good coverage of natural distributions, which is conducive to the sharing of model knowledge. To further tackle the privacy issue, we exploit the lightweight additive secret sharing technique to construct a series of tailored cryptographic protocols for key building blocks such as secure prediction. Our protocols implement ciphertext operations through simple vectorized computations, which are friendly with GPUs and can be processed by highly-optimized CUDA kernels. Extensive evaluations demonstrate that PrivHFL outperforms prior art up to two orders of magnitude in efficiency and realizes significant accuracy gains on top of the stand-alone method.

****

ViTGAN: Training GANs with Vision Transformers

Kwonjoon Lee,Huiwen Chang,Lu Jiang,Han Zhang,Zhuowen Tu,Ce Liu

Recently, Vision Transformers (ViTs) have shown competitive performance on image recognition while requiring less vision-specific inductive biases. In this paper, we investigate if such performance can be extended to image generation. To this end, we integrate the ViT architecture into generative adversarial networks (GANs). For ViT discriminators, we observe that existing regularization methods for GANs interact poorly with self-attention, causing serious instability during training. To resolve this issue, we introduce several novel regularization techniques for training GANs with ViTs. For ViT generators, we examine architectural choices for latent and pixel mapping layers to faciliate convergence. Empirically, our approach, named ViTGAN, achieves comparable performance to the leading CNN- based GAN models on three datasets: CIFAR-10, CelebA, and LSUN bedroom.

****

Should we Replace CNNs with Transformers for Medical Images?

Christos Matsoukas,Johan Fredin Haslum,Moein Sorkhei,Magnus Soderberg,Kevin Smith

Convolutional Neural Networks (CNNs) have reigned for a decade as the de facto approach to automated medical image diagnosis, pushing the state-of-the-art in classification, detection and segmentation tasks. Recently, vision transformers (ViTs) have appeared as a competitive alternative to CNNs, yielding impressive levels of performance in the natural image domain, while possessing several interesting properties that could prove beneficial for medical imaging tasks. In this work, we explore whether it is feasible to switch to transformer-based models in the medical imaging domain as well, or if we should keep working with CNNs - can we trivially replace CNNs with transformers? We consider this question in a series of experiments on several standard medical image benchmark datasets and tasks. Our findings show that, while CNNs perform better if trained from scratch, off-the-shelf vision transformers are on par with CNNs when pretrained on ImageNet in both classification and segmentation tasks. Further, ViTs often outperform their CNN counterparts when pretrained using self-supervision.

****

Bit-aware Randomized Response for Local Differential Privacy in Federated Learning

Phung Lai,Hai Phan,Li Xiong,Khang Phuc Tran,My Thai,Tong Sun,Franck Dernoncourt,Jiuxiang Gu,Nikolaos Barmpalios,Rajiv Jain

In this paper, we develop BitRand, a bit-aware randomized response algorithm, to preserve local differential privacy (LDP) in federated learning (FL). We encode embedded features extracted from clients' local data into binary encoding bits, in which different bits have different impacts on the embedded features. Based upon that, we randomize all the bits to preserve LDP with three key advantages: (1) Bit-aware: Bits with a more substantial influence on the model utility have smaller randomization probabilities, and vice-versa, under the same privacy protection; (2) Dimension-elastic: Increasing the dimensions of embedded features, gradients, model outcomes, and training rounds marginally affect the randomization probabilities of binary encoding bits under the same privacy protection; and (3) LDP protection is achieved for both embedded features and labels with tight privacy loss and expected error bounds ensuring high model utility. Extensive theoretical and experimental results show that our BitRand significantly outperforms various baseline approaches in text and image classification.

****

Distributionally Robust Fair Principal Components via Geodesic Descents

Hieu Vu,Toan Tran,Man-Chung Yue,Viet Anh Nguyen

Principal component analysis is a simple yet useful dimensionality reduction technique in modern machine learning pipelines. In consequential domains such as college admission, healthcare and credit approval, it is imperative to take into account emerging criteria such as the fairness and the robustness of the learned projection. In this paper, we propose a distributionally robust optimization problem for principal component analysis which internalizes a fairness criterion in the objective function. The learned projection thus balances the trade-off between the total reconstruction error and the reconstruction error gap between subgroups, taken in the min-max sense over all distributions in a moment-based ambiguity set. The resulting optimization problem over the Stiefel manifold can be efficiently solved by a Riemannian subgradient descent algorithm with a sub-linear convergence rate. Our experimental results on real-world datasets show the merits of our proposed method over state-of-the-art baselines.

****

Self-Distribution Distillation: Efficient Uncertainty Estimation

Yassir Fathullah,Mark Gales

Deep learning is increasingly being applied in safety-critical domains. For these scenarios it is important to know the level of uncertainty in a model’s prediction to ensure that appropriate decisions are made by a system. Deep ensembles are the de-facto standard approach to obtaining various measures of uncertainty. However, ensembles normally significantly increase the resources required in both the training and deployment phases. Approaches have been developed that typically address the costs in one of these phases. In this work we propose a novel training approach, self-distribution distillation (S2D), which is able to efficiently, both in time and memory, train a single model that can estimate uncertainties in an integrated training phase. Furthermore it is possible to build ensembles of these models and apply ensemble distillation approaches, hierarchical distribution distillation, in cases where one is less limited by computational resources in the training phase, but still requires efficiency in the deployment phase. Experiments on CIFAR-100 showed that S2D models outperformed standard models and Monte-Carlo dropout. Additional out-of-distribution detection experiments on LSUN, Tiny ImageNet, SVHN showed that even a standard deep ensemble can be outperformed using S2D based ensembles and novel distilled models.

****

Addressing the Stability-Plasticity Dilemma via Knowledge-Aware Continual Learning

Ghada Sokar,Decebal Constantin Mocanu,Mykola Pechenizkiy

Continual learning agents should incrementally learn a sequence of tasks while s atisfying two main desiderata: accumulating on previous knowledge without forgetting and transferring previous relevant knowledge to help in future learning. Existing research largely focuses on alleviating the catastrophic forgetting problem. There, an agent is altered to prevent forgetting based solely on previous tasks. This hinders the balance between preventing forgetting and maximizing the forward transfer. In response to this, we investigate the stability-plasticity dilemma to determine which model components are eligible to be reused, added, fixed, or updated to achieve this balance. We address the class incremental learning scenario where the agent is prone to ambiguities between old and new classes. With our proposed Knowledge-Aware contiNual learner (KAN), we demonstrate that considering the semantic similarity between old and new classes helps in achieving this balance. We show that being aware of existing knowledge helps in: (1) increasing the forward transfer from similar knowledge, (2) reducing the required capacity by leveraging existing knowledge, (3) protecting dissimilar knowledge, and (4) increasing robustness to the class order in the sequence. We evaluated sequences of similar tasks, dissimilar tasks, and a mix of both constructed from the two commonly used benchmarks for class-incremental learning; CIFAR-10 and CIFAR-100.

****

Continual Learning in Deep Networks: an Analysis of the Last Layer

Timothee LESORT,Thomas George,Irina Rish

We study how different output layers in a deep neural network learn and forget in continual learning settings. The following three factors can affect catastrophic forgetting in the output layer: (1) weights modifications, (2) interference, and (3) projection drift. In this paper, our goal is to provide more insights into how changing the output layers may address (1) and (2). Some potential solutions to those issues are proposed and evaluated here in several continual learning scenarios. We show that the best-performing type of the output layer depends on the data distribution drifts and/or the amount of data available. In particular, in some cases where a standard linear layer would fail, it turns out that changing parameterization is sufficient in order to achieve a significantly better performance, whithout introducing a continual-learning algorithm and instead using the standard SGD to train a model. Our analysis and results shed light on the dynamics of the output layer in continual learning scenarios, and suggest a way of selecting the best type of output layer for a given scenario.

****

Multi-Task Neural Processes

Jiayi Shen,Xiantong Zhen,Marcel Worring,Ling Shao

Neural processes have recently emerged as a class of powerful neural latent variable models that combine the strengths of neural networks and stochastic processes. As they can encode contextual data in the network's function space, they offer a new way to model task relatedness in multi-task learning. To study its potential, we develop multi-task neural processes, a new variant of neural processes for multi-task learning. In particular, we propose to explore transferable knowledge from related tasks in the function space to provide inductive bias for improving each individual task. To do so, we derive the function priors in a hierarchical Bayesian inference framework, which enables each task to incorporate the shared knowledge provided by related tasks into its context of the prediction function. Our multi-task neural processes methodologically expand the scope of vanilla neural processes and provide a new way of exploring task relatedness in function spaces for multi-task learning. The proposed multi-task neural processes are capable of learning multiple tasks with limited labeled data and in the presence of domain shift. We perform extensive experimental evaluations on several benchmarks for the multi-task regression and classification tasks. The results demonstrate the effectiveness of multi-task neural processes in transferring useful knowledge among tasks for multi-task learning and superior performance in multi-task classification and brain image segmentation.

****

Dataset transformations trade-offs to adapt machine learning methods across domains

Napoleon Costilla-Enriquez,Yang Weng

Machine learning-based methods have been proved to be quite successful in different domains. However, applying the same techniques across disciplines is not a trivial task with benefits and drawbacks. In the literature, the most common approach is to convert a dataset into the same format as the original domain to employ the same architecture that was successful in the original domain. Although this approach is fast and convenient, we argue it is suboptimal due to the lack of tailoring to the specific problem at hand. To prove our point, we examine dataset transformations used in the literature to adapt machine learning-based methods across domains and show that these dataset transformations are not always beneficial in terms of performance. In addition, we show that these data transformations open the door to unforeseen vulnerabilities in the new applied different domain. To quantify how different the original dataset is with respect to the transformed one, we compute the dataset distances via Optimal Transport. Also, we present simulations with the original and transformed data to show that the data conversion is not always needed and exposes the new domain to unsought menaces.

****

Unit Ball Model for Embedding Hierarchical Structures in the Complex Hyperbolic Space

Huiru Xiao,Caigao JIANG,Yangqiu Song,james Y zhang,Junwu Xiong

Learning the representation of data with hierarchical structures in the hyperbolic space attracts increasing attention in recent years. Due to the constant negative curvature, the hyperbolic space resembles tree metrics and captures the tree-like properties naturally, which enables the hyperbolic embeddings to improve over traditional Euclidean models. However, most real-world hierarchically structured data such as taxonomies and multitree networks have varying local structures and they are not trees, thus they do not ubiquitously match the constant curvature property of the hyperbolic space. To address this limitation of hyperbolic embeddings, we explore the complex hyperbolic space, which has the variable negative curvature, for representation learning. Specifically, we propose to learn the embeddings of hierarchically structured data in the unit ball model of the complex hyperbolic space. The unit ball model based embeddings have a more powerful representation capacity to capture a variety of hierarchical structures. Through experiments on synthetic and real-world data, we show that our approach improves over the hyperbolic embedding models significantly. We also explore the competence of complex hyperbolic geometry on the multitree structure and 1-N structure.

****

Transformer with a Mixture of Gaussian Keys

Tam Minh Nguyen,Tan Minh Nguyen,Dung Duy Le,Nguyen Duy Khuong,Viet Anh TRAN,Richard Baraniuk,Nhat Ho,Stanley Osher

Multi-head attention is a driving force behind state-of-the-art transformers which achieve remarkable performance across a variety of natural language processing (NLP) and computer vision tasks. It has been observed that for many applications, those attention heads learn redundant embedding, and most of them can be removed without degrading the performance of the model. Inspired by this observation, we propose Transformer with a Mixture of Gaussian Keys (Transformer-MGK), a novel transformer architecture that replaces redundant heads in transformers with a mixture of keys at each head. These mixtures of keys follow a Gaussian mixture model and allow each attention head to focus on different parts of the input sequence efficiently. Compared to its conventional transformer counterpart, Transformer-MGK accelerates training and inference, has fewer parameters, and requires fewer FLOPs to compute while achieving comparable or better accuracy across tasks. Transformer-MGK can also be easily extended to use with linear attention. We empirically demonstrate the advantage of Transformer-MGK in a range of practical applications including language modeling and tasks that involve very long sequences. On the Wikitext-103 and Long Range Arena benchmark, Transformer-MGKs with 4 heads attain comparable or better performance to the baseline transformers with 8 heads.

****

StyleAlign: Analysis and Applications of Aligned StyleGAN Models

Zongze Wu,Yotam Nitzan,Eli Shechtman,Dani Lischinski

In this paper, we perform an in-depth study of the properties and applications of aligned generative models.

We refer to two models as aligned if they share the same architecture, and one of them (the child) is obtained from the other (the parent) via fine-tuning to another domain, a common practice in transfer learning. Several works already utilize some basic properties of aligned StyleGAN models to perform image-to-image translation. Here, we perform the first detailed exploration of model alignment, also focusing on StyleGAN. First, we empirically analyze aligned models and provide answers to important questions regarding their nature. In particular, we find that the child model's latent spaces are semantically aligned with those of the parent, inheriting incredibly rich semantics, even for distant data domains such as human faces and churches. Second, equipped with this better understanding, we leverage aligned models to solve a diverse set of tasks. In addition to image translation, we demonstrate fully automatic cross-domain image morphing. We further show that zero-shot vision tasks may be performed in the child domain, while relying exclusively on supervision in the parent domain. We demonstrate qualitatively and quantitatively that our approach yields state-of-the-art results, while requiring only simple fine-tuning and inversion.

****

A Collaborative Attention Adaptive Network for Financial Market Forecasting

Qiuyue Zhang,Yunfeng Zhang,Fangxun Bao,Caiming Zhang,Peide Liu,Xunxiang Yao

Forecasting the financial market with social media data and real market prices is a valuable issue for market participants, which helps traders make more appropriate trading decisions. However, taking into account the differences of different data types, how to use a fusion method adapted to financial data to fuse real market prices and tweets from social media, so that the prediction model can fully integrate different types of data remains a challenging problem. To address these problems, we propose a collaborative attention adaptive Transformer approach to financial market forecasting (CAFF), including parallel extraction of tweets and price features, parameter-level fusion and a joint feature processing module, that can successfully deeply fuse tweets and real prices in view of the fusion method. Extensive experimentation is performed on tweets and historical price of stock market, our method can achieve a better accuracy compared with the state-of-the-art methods on two evaluation metrics. Moreover, tweets play a relatively more critical role in the CAFF framework. Additional stock trading simulations show that an actual trading strategy based on our proposed model can increase profits; thus, the model has practical application value.

****

Wakening Past Concepts without Past Data: Class-incremental Learning from Placebos

Yaoyao Liu,Bernt Schiele,Qianru Sun

Not forgetting knowledge about previous classes is one of the key challenges in class-incremental learning (CIL). A common technique to address this challenge is knowledge distillation (KD) that penalizes inconsistencies across models of subsequent phases. As old-class data is scarce, the KD loss mainly uses new class data. However, we empirically observe that this both harms learning of new classes and also underperforms to distil old class knowledge from the previous phase model. To address this issue, we propose to compute the KD loss using placebo data chosen from a free image stream (e.g., Google Images), which is both simple and surprisingly effective even when there is no class overlap between the placebos and the old data. When the image stream is available, we use an evaluation function to quickly judge the quality of candidate images (good or bad placebos) and collect good ones. For training this function, we sample pseudo CIL tasks from the data in the 0-th phase and design a reinforcement learning algorithm. Our method does not require any additional supervision or memory budget, and can significantly improve a number of top-performing CIL methods, in particular on higher-resolution benchmarks, e.g., ImageNet-1k and ImageNet-Subset, and with a lower memory budget for old class exemplars, e.g., five exemplars per class.

****

Symmetric Machine Theory of Mind
Melanie Sclar,Graham Neubig,Yonatan Bisk
Theory of mind (ToM), the ability to understand others' thoughts and desires, is a cornerstone of human intelligence. Because of this, a number of previous works have attempted to measure the ability of machines to develop a theory of mind, with one agent attempting to understand anothers' internal "mental state''. However, ToM agents are often tested as passive observers or in tasks with specific predefined roles, such as speaker-listener scenarios. In this work, we propose to model machine theory of mind in a more flexible and symmetric scenario; a multi-agent environment SymmToM where all agents can speak, listen, see other agents, and move freely through a grid world. An effective strategy to solve SymmToM requires developing theory of mind to maximize each agent's rewards. We show that multi-agent deep-reinforcement learning models that model the mental states of other agents achieve significant performance improvements over agents with no such ToM model. At the same time, our best agents fail to achieve performance comparable to agents with access to the gold-standard mental state of other agents, demonstrating that the modeling of theory of mind in multi-agent scenarios is very much an open challenge.

****

Understanding and Improving Graph Injection Attack by Promoting Unnoticeability
Yongqiang Chen,Han Yang,Yonggang Zhang,MA KAILI,Tongliang Liu,Bo Han,James Cheng
Recently Graph Injection Attack (GIA) emerges as a practical attack scenario on Graph Neural Networks (GNNs), where the adversary can merely inject few malicious nodes instead of modifying existing nodes or edges, i.e., Graph Modification Attack (GMA). Although GIA has achieved promising results, little is known about why it is successful and whether there is any pitfall behind the success. To understand the power of GIA, we compare it with GMA and find that GIA can be provably more harmful than GMA due to its relatively high flexibility. However, the high flexibility will also lead to great damage to the homophily distribution of the original graph, i.e., similarity among neighbors. Consequently, the threats of GIA can be easily alleviated or even prevented by homophily-based defenses designed to recover the original homophily. To mitigate the issue, we introduce a novel constraint – homophily unnoticeability that enforces GIA to preserve the homophily, and propose Harmonious Adversarial Objective (HAO) to instantiate it. Extensive experiments verify that GIA with HAO can break homophily-based defenses and outperform previous GIA attacks by a significant margin. We believe our methods can serve for a more reliable evaluation of the robustness of GNNs.

****

Learning Neural Implicit Functions as Object Representations for Robotic Manipulation
Jung-Su Ha,Danny Driess,Marc Toussaint
Robotic manipulation planning is the problem of finding a sequence of robot configurations that involves interactions with objects in the scene, e.g., grasp, placement, tool-use, etc. To achieve such interactions, traditional approaches require hand-designed features and object representations, and it still remains an open question how to describe such interactions with arbitrary objects in a flexible and efficient way. Inspired by neural implicit representations in 3D modeling, e.g. NeRF, we propose a method to represent objects as neural implicit functions upon which we can define and jointly train interaction features. The proposed pixel-aligned representation is directly inferred from camera images with known camera geometry, naturally acting as a perception component in the whole manipulation pipeline, while at the same time enabling sequential robot manipulation planning.

****

Spatio-temporal Disentangled representation learning for mobility prediction
Sichen Zhao,Wei Shao,Jeffrey Chan,Flora D. Salim
Spatio-temporal (ST) prediction task like mobility forecasting is of great significance to traffic management and public safety.

There is an increasing number of works proposed for mobility forecasting problems recently, and they typically focus on better extraction of the features from the spatial and temporal domains. Although prior works show promising results on more accurate predictions, they still suffer in characterising and separating the dynamic and static components, making it difficult to make further improvements. Disentangled representation learning separates the learnt latent representation into independent variables associated with semantic factors. It offers a better separation of the spatial and temporal features, which could improve the performance of mobility forecasting models. In this work, we propose a VAE-based architecture for learning the disentangled representation from real spatio-temporal data for mobility forecasting. Our deep generative model learns a latent representation that (i) separates the temporal dynamics of the data from the spatially varying component and generates effective reconstructions; (ii) is able to achieve state-of-the-art performance across multiple spatio-temporal datasets. Moreover, we investigate the effectiveness of our method by eliminating the non-informative features from the learnt representations, and the results show that models can benefit from this operation.

****

Double Descent in Adversarial Training: An Implicit Label Noise Perspective
Chengyu Dong,Liyuan Liu,Jingbo Shang

Here, we show that the robust overfitting shall be viewed as the early part of an epoch-wise double descent --- the robust test error will start to decrease again after training the model for a considerable number of epochs. Inspired by our observations, we further advance the analyses of double descent to understand robust overfitting better. In standard training, double descent has been shown to be a result of label flipping noise. However, this reasoning is not applicable in our setting, since adversarial perturbations are believed not to change the label. Going beyond label flipping noise, we propose to measure the mismatch between the assigned and (unknown) true label distributions, denoted as \emph{implicit label noise}. We show that the traditional labeling of adversarial examples inherited from their clean counterparts will lead to implicit label noise. Towards better labeling, we show that predicted distribution from a classifier, after scaling and interpolation, can provably reduce the implicit label noise under mild assumptions. In light of our analyses, we tailored the training objective accordingly to effectively mitigate the double descent and verified its effectiveness on three benchmark datasets.

****

How Does the Task Landscape Affect MAML Performance?
Liam Collins,Aryan Mokhtari,Sanjay Shakkottai

Model-Agnostic Meta-Learning (MAML) has become increasingly popular for training models that can quickly adapt to new tasks via one or few stochastic gradient descent steps. However, the MAML objective is significantly more difficult to optimize compared to standard non-adaptive learning (NAL), and little is understood about how much MAML improves over NAL in terms of the fast adaptability of their solutions in various scenarios. We analytically address this issue in a linear regression setting consisting of a mixture of easy and hard tasks, where hardness is related to the rate that gradient descent converges on the task. Specifically, we prove that in order for MAML to achieve substantial gain over NAL, (i) there must be some discrepancy in hardness among the tasks, and (ii) the optimal solutions of the hard tasks must be closely packed with the center far from the center of the easy tasks optimal solutions. We also give numerical and analytical results suggesting that these insights apply to two-layer neural networks. Finally, we provide few-shot image classification experiments that support our insights for when MAML should be used and emphasize the importance of training MAML on hard tasks in practice.

****

Data Quality Matters For Adversarial Training: An Empirical Study
Chengyu Dong,Liyuan Liu,Jingbo Shang

Multiple intriguing problems are hovering in adversarial training, including rob ust overfitting, robustness overestimation, and robustness-accuracy trade-off. These problems pose great challenges to both reliable evaluation and practical deployment. Here, we empirically show that these problems share one common cause --- low-quality samples in the dataset. Specifically, we first propose a strategy to measure the data quality based on the learning behaviors of the data during adversarial training and find that low-quality data may not be useful and even detrimental to the adversarial robustness. We then design controlled experiments to investigate the interconnections between data quality and problems in adversarial training. We find that when low-quality data is removed, robust overfitting and robustness overestimation can be largely alleviated; and robustness-accuracy trade-off becomes less significant. These observations not only verify our intuition about data quality but may also open new opportunities to advance adversarial training.

****

Interactively Generating Explanations for Transformer Language Models
Patrick Schramowski,Felix Friedrich,Christopher Tauchmann,Kristian Kersting
Transformer language models are state-of-the-art in a multitude of NLP tasks. Despite these successes, their opaqueness remains problematic. Recent methods aiming to provide interpretability and explainability to black-box models primarily focus on post-hoc explanations of (sometimes spurious) input-output correlations. Instead, we emphasize using prototype networks directly incorporated into the model architecture and hence explain the reasoning process behind the network's decisions. Moreover, while our architecture performs on par with several language models, it enables one to learn from user interactions. This not only offers a better understanding of language models but uses human capabilities to incorporate knowledge outside of the rigid range of purely data-driven approaches.

****

Physics-Informed Neural Operator for Learning Partial Differential Equations
Zongyi Li,Hongkai Zheng,Nikola Borislavov Kovachki,David Jin,Haoxuan Chen,Burigede Liu,Andrew Stuart,Kamyar Azizzadenesheli,Anima Anandkumar
Machine learning methods have recently shown promise in solving partial differential equations (PDEs). They can be classified into two broad categories: solution function approximation, and operator learning. The Physics-Informed Neural Network (PINN) is an example of the former while the Fourier neural operator (FNO) is an example of the latter. Both these approaches have shortcomings. The optimization in PINN is challenging and prone to failure, especially on multi-scale dynamic systems. FNO does not suffer from this optimization issue since it carries out supervised learning on a given dataset, but obtaining such data may be too expensive or infeasible. In this work, we propose the physics-informed neural operator (PINO), where we combine the operating-learning and function-optimization frameworks, and this improves convergence rates and accuracy over both PINN and FNO models. In the operator-learning phase, PINO learns the solution operator over multiple instances of the parametric PDE family. In the test-time optimization phase, PINO optimizes the pre-trained operator ansatz for the querying instance of the PDE. Experiments show PINO outperforms previous ML methods on many popular PDE families while retaining the extraordinary speed-up of FNO compared to solvers. In particular, PINO accurately solves long temporal transient flows and chaotic Kolmogorov flows, while PINN and other methods fail to converge to a reasonable accuracy.

****

Learning to Guide and to be Guided in the Architect-Builder Problem
Paul Barde,Tristan Karch,Derek Nowrouzezahrai,Clément Moulin-Frier,Christopher Pal,Pierre-Yves Oudeyer
We are interested in interactive agents that learn to coordinate, namely, a $builder$ -- which performs actions but ignores the goal of the task, i.e. has no access to rewards -- and an $architect$ which guides the builder towards the goal of the task.

We define and explore a formal setting where artificial agents are equipped with mechanisms that allow them to simultaneously learn a task while at the same time evolving a shared communication protocol.

Ideally, such learning should only rely on high-level communication priors and be able to handle a large variety of tasks and meanings while deriving communication protocols that can be reused across tasks.

The field of Experimental Semiotics has shown the extent of human proficiency at learning from a priori unknown instructions meanings. Therefore, we take inspiration from it and present the Architect-Builder Problem (ABP): an asymmetrical setting in which an architect must learn to guide a builder towards constructing a specific structure. The architect knows the target structure but cannot act in the environment and can only send arbitrary messages to the builder. The builder on the other hand can act in the environment, but receives no rewards nor has any knowledge about the task, and must learn to solve it relying only on the messages sent by the architect. Crucially, the meaning of messages is initially not defined nor shared between the agents but must be negotiated throughout learning.

Under these constraints, we propose Architect-Builder Iterated Guiding (ABIG), a solution to the Architect-Builder Problem where the architect leverages a learned model of the builder to guide it while the builder uses self-imitation learning to reinforce its guided behavior. To palliate to the non-stationarity induced by the two agents concurrently learning, ABIG structures the sequence of interactions between the agents into interaction frames. We analyze the key learning mechanisms of ABIG and test it in a 2-dimensional instantiation of the ABP where tasks involve grasping cubes, placing them at a given location, or building various shapes. In this environment, ABIG results in a low-level, high-frequency, guiding communication protocol that not only enables an architect-builder pair to solve the task at hand, but that can also generalize to unseen tasks.

****

Test-Time Adaptation to Distribution Shifts by Confidence Maximization and Input Transformation

Chaithanya Kumar Mummadi,Robin Hutmacher,Kilian Rambach,Evgeny Levinkov,Thomas Brox,Jan Hendrik Metzen

Deep neural networks often exhibit poor performance on data that is unlikely under the train-time data distribution, for instance data affected by corruptions. Previous works demonstrate that test-time adaptation to data shift, for instance using entropy minimization, effectively improves performance on such shifted distributions. This paper focuses on the fully test-time adaptation setting, where only unlabeled data from the target distribution is required. This allows adapting arbitrary pretrained networks. Specifically, we propose a novel loss that improves test-time adaptation by addressing both premature convergence and instability of entropy minimization. This is achieved by replacing the entropy by a non-saturating surrogate and adding a diversity regularizer based on batch-wise entropy maximization that prevents convergence to trivial collapsed solutions. Moreover, we propose to prepend an input transformation module to the network that can partially undo test-time distribution shifts. Surprisingly, this preprocessing can be learned solely using the fully test-time adaptation loss in an end-to-end fashion without any target domain labels or source domain data. We show that our approach outperforms previous work in improving the robustness of publicly available pretrained image classifiers to common corruptions on such challenging benchmarks as ImageNet-C.

****

Phase Collapse in Neural Networks

Florentin Guth,John Zarka,Stéphane Mallat

Deep convolutional classifiers linearly separate image classes and improve accuracy as depth increases. They progressively reduce the spatial dimension whereas the number of channels grows with depth. Spatial variability is therefore transformed into variability along channels. A fundamental challenge is to understand the role of non-linearities together with convolutional filters in this transformation. ReLUs with biases are often interpreted as thresholding operators that improve discrimination through sparsity. This paper demonstrates that it is a different mechanism called \emph{phase collapse} which eliminates spatial variability while linearly separating classes. We show that collapsing the phases of comp lex wavelet coefficients is sufficient to reach the classification accuracy of ResNets of similar depths. However, replacing the phase collapses with thresholding operators that enforce sparsity considerably degrades the performance. We explain these numerical results by showing that the iteration of phase collapses progressively improves separation of classes, as opposed to thresholding non-linearities.

****

SPIRAL: Self-supervised Perturbation-Invariant Representation Learning for Speech Pre-Training

Wenyong Huang,Zhenhe Zhang,Yu Ting Yeung,Xin Jiang,Qun Liu

We introduce a new approach for speech pre-training named SPIRAL which works by learning denoising representation of perturbed data in a teacher-student framework.

Specifically, given a speech utterance, we first feed the utterance to a teacher network to obtain corresponding representation. Then the same utterance is perturbed and fed to a student network. The student network is trained to output representation resembling that of the teacher. At the same time, the teacher network is updated as moving average of student's weights over training steps. In order to prevent representation collapse, we apply an in-utterance contrastive loss as pre-training objective and impose position randomization on the input to the teacher. SPIRAL achieves competitive or better results compared to state-of-the-art speech pre-training method wav2vec 2.0, with significant reduction of training cost (80% for BASE model, 65% for LARGE model).

Furthermore, we address the problem of noise-robustness that is critical to real-world speech applications. We propose multi-condition pre-training by perturbing the student's input with various types of additive noise. We demonstrate that multi-condition pre-trained SPIRAL models are more robust to noisy speech (9.0% - 13.3% relative word error rate reduction on real noisy test data), compared to applying multi-condition training solely in the fine-tuning stage. Source code is available at https://github.com/huawei-noah/Speech-Backbones/tree/main/SPIRAL.

****

Improving the Accuracy of Learning Example Weights for Imbalance Classification

Yuqi Liu,Bin Cao,Jing Fan

To solve the imbalance classification, methods of weighting examples have been proposed. Recent work has studied to assign adaptive weights to training examples through learning mechanisms, that is, the weights, similar to classification models, are regarded as parameters that need to be learned. However, the algorithms in recent work use local information to approximately optimize the weights, which may lead to inaccurate learning of the weights. In this work, we first propose a novel mechanism of learning with a constraint, which can accurately train the weights and model. Then, we propose a combined method of our learning mechanism and the work by Hu et al., which can promote each other to perform better. Our proposed method can be applied to any type of deep network model. Experiments show that compared with the state-of-the-art algorithms, our method has significant improvement in varieties of settings, including text and image classification over different imbalance ratios, binary and multi-class classification.

****

Predictive Maintenance for Optical Networks in Robust Collaborative Learning

Khouloud Abdelli,JOO YEON CHO

Machine learning (ML) has recently emerged as a powerful tool to enhance the proactive optical network maintenance and thereby, improve network reliability and operational efficiency, and reduce unplanned downtime and maintenance costs. However, it is challenging to develop an accurate and reliable ML based prognostic models due mainly to the unavailability of sufficient amount of training data since the device failure does not occur often in optical networks. Federated learning (FL) is a promising candidate to tackle the aforementioned challenge by enabling the development of a global ML model using datasets owned by many vendors without revealing their business-confidential data. While FL greatly enhances the data privacy, a global model can be strongly affected by a malicious local mod el. We propose a robust collaborative learning framework for predictive maintenance on cross-vendor in a dishonest setting. Our experiments confirm that a global ML model can be accurately built with sensitive datasets in federated learning even when a subset of vendors behave dishonestly.

****

POETREE: Interpretable Policy Learning with Adaptive Decision Trees

Alizée Pace,Alex Chan,Mihaela van der Schaar

Building models of human decision-making from observed behaviour is critical to better understand, diagnose and support real-world policies such as clinical care. As established policy learning approaches remain focused on imitation performance, they fall short of explaining the demonstrated decision-making process. Policy Extraction through decision Trees (POETREE) is a novel framework for interpretable policy learning, compatible with fully-offline and partially-observable clinical decision environments -- and builds probabilistic tree policies determining physician actions based on patients' observations and medical history. Fully-differentiable tree architectures are grown incrementally during optimization to adapt their complexity to the modelling task, and learn a representation of patient history through recurrence, resulting in decision tree policies that adapt over time with patient information. This policy learning method outperforms the state-of-the-art on real and synthetic medical datasets, both in terms of understanding, quantifying and evaluating observed behaviour as well as in accurately replicating it -- with potential to improve future decision support systems.

****

How memory architecture affects learning in a simple POMDP: the two-hypothesis testing problem

Mario Geiger,Christophe Eloy,Matthieu Wyart

Reinforcement learning is generally difficult for partially observable Markov decision processes (POMDPs), which occurs when the agent's observation is partial or noisy. To seek good performance in POMDPs, one strategy is to endow the agent with a finite memory, whose update is governed by the policy. However, policy optimization is non-convex in that case and can lead to poor training performance for random initialization. The performance can be empirically improved by constraining the memory architecture, then sacrificing optimality to facilitate training. Here we study this trade-off in a two-hypothesis testing problem, akin to the two-arm bandit problem. We compare two extreme cases: (i) the random access memory where any transitions between M memory states are allowed and (ii) a fixed memory where the agent can access its last m actions and rewards. For (i), the probability q to play the worst arm is known to be exponentially small in M for the optimal policy. Our main result is to show that similar performance can be reached for (ii) as well, despite the simplicity of the memory architecture: using a conjecture on Gray-ordered binary necklaces, we find policies for which q is exponentially small in 2^m, i.e. $q\sim\alpha^{2^m}$ with $\alpha < 1$. In addition, we observe empirically that training from random initialization leads to very poor results for (i), and significantly better results for (ii) thanks to the constraints on the memory architecture.

****

Empirical Study of the Decision Region and Robustness in Deep Neural Networks

Seongjin Park,Haedong Jeong,Giyoung Jeon,Jaesik Choi

In general, the Deep Neural Networks (DNNs) is evaluated by the generalization performance measured on the unseen data excluded from the training phase. Along with the development of DNNs, the generalization performance converges to the state-of-the-art and it becomes difficult to evaluate DNNs solely based on the generalization performance. The robustness against the adversarial attack has been used as an additional metric to evaluate DNNs by measuring the vulnerability of them. However, few researches have been performed to analyze the adversarial robustness in terms of the geometry in DNNs. In this work, we perform empirical study to analyze the internal properties of DNNs which affect model robustness under adversarial attacks. Especially, we propose the novel concept Populated Region Set (PRS) where train samples populated more frequently to represent the interna l properties of DNNs in the practical setting. From the systematic experiments with the proposed concept, we provide empirical evidences to validate that the low PRS ratio has strong relationship with the adversarial robustness of DNNs.

****

Generating Videos with Dynamics-aware Implicit Generative Adversarial Networks

Sihyun Yu,Jihoon Tack,Sangwoo Mo,Hyunsu Kim,Junho Kim,Jung-Woo Ha,Jinwoo Shin

In the deep learning era, long video generation of high-quality still remains challenging due to the spatio-temporal complexity and continuity of videos. Existing prior works have attempted to model video distribution by representing videos as 3D grids of RGB values, which impedes the scale of generated videos and neglects continuous dynamics. In this paper, we found that the recent emerging paradigm of implicit neural representations (INRs) that encodes a continuous signal into a parameterized neural network effectively mitigates the issue. By utilizing INRs of video, we propose dynamics-aware implicit generative adversarial network (DIGAN), a novel generative adversarial network for video generation. Specifically, we introduce (a) an INR-based video generator that improves the motion dynamics by manipulating the space and time coordinates differently and (b) a motion discriminator that efficiently identifies the unnatural motions without observing the entire long frame sequences. We demonstrate the superiority of DIGAN under various datasets, along with multiple intriguing properties, e.g., long video synthesis, video extrapolation, and non-autoregressive video generation. For example, DIGAN improves the previous state-of-the-art FVD score on UCF-101 by 30.7% and can be trained on 128 frame videos of 128x128 resolution, 80 frames longer than the 48 frames of the previous state-of-the-art method.

****

Efficient Learning of Safe Driving Policy via Human-AI Copilot Optimization

Quanyi Li,Zhenghao Peng,Bolei Zhou

Human intervention is an effective way to inject human knowledge into the training loop of reinforcement learning, which can bring fast learning and ensured training safety. Given the very limited budget of human intervention, it remains challenging to design when and how human expert interacts with the learning agent in the training. In this work, we develop a novel human-in-the-loop learning method called Human-AI Copilot Optimization (HACO).To allow the agent's sufficient exploration in the risky environments while ensuring the training safety, the human expert can take over the control and demonstrate how to avoid probably dangerous situations or trivial behaviors. The proposed HACO then effectively utilizes the data both from the trial-and-error exploration and human's partial demonstration to train a high-performing agent. HACO extracts proxy state-action values from partial human demonstration and optimizes the agent to improve the proxy values meanwhile reduce the human interventions. The experiments show that HACO achieves a substantially high sample efficiency in the safe driving benchmark. HACO can train agents to drive in unseen traffic scenarios with a handful of human intervention budget and achieve high safety and generalizability, outperforming both reinforcement learning and imitation learning baselines with a large margin. Code and demo video are included in the supplementary materials.

****

Detecting Adversarial Examples Is (Nearly) As Hard As Classifying Them

Florian Tramer

Making classifiers robust to adversarial examples is challenging.
Thus, many defenses tackle the seemingly easier task of \emph{detecting} perturbed inputs.

We show a barrier towards this goal. We prove a general \emph{hardness reduction} between detection and classification of adversarial examples: given a robust detector for attacks at distance ϵ (in some metric), we show how to build a similarly robust (but inefficient) \emph{classifier} for attacks at distance $\epsilon/2$---and vice-versa.

Our reduction is computationally inefficient, and thus cannot be used to build practical classifiers. Instead, it is a useful sanity check to test whether empir ical detection results imply something much stronger than the authors presumably anticipated.

To illustrate, we revisit 14 empirical detector defenses published over the past years. For $12/14$ defenses, we show that the claimed detection results imply an inefficient classifier with robustness far beyond the state-of-the-art--- thus casting some doubts on the results' validity.

Finally, we show that our reduction applies in both directions: a robust classifier for attacks at distance $\epsilon/2$ implies an inefficient robust detector at distance ϵ. Thus, we argue that robust classification and robust detection should be regarded as (near)-equivalent problems.
****

Re-evaluating Word Mover's Distance

Ryoma Sato,Makoto Yamada,Hisashi Kashima

The word mover's distance (WMD) is a fundamental technique for measuring the similarity of two documents. As the crux of WMD, it can take advantage of the underlying geometry of the word space by employing an optimal transport formulation. The original study on WMD reported that WMD outperforms classical baselines such as bag-of-words (BOW) and TF-IDF by significant margins in various datasets. In this paper, we point out that the evaluation in the original study could be misleading. We re-evaluate the performances of WMD and the classical baselines and find that the classical baselines are competitive with WMD if we employ an appropriate preprocessing, i.e., L1 normalization. In addition, We introduce an analogy between WMD and L1-normalized BOW and find that not only the performance of WMD but also the distance values resemble those of BOW in high dimensional spaces.
****

Enhancing Cross-lingual Transfer by Manifold Mixup

Huiyun Yang,Huadong Chen,Hao Zhou,Lei Li

Based on large-scale pre-trained multilingual representations, recent cross-lingual transfer methods have achieved impressive transfer performances. However, the performance of target languages still lags far behind the source language. In this paper, our analyses indicate such a performance gap is strongly associated with the cross-lingual representation discrepancy. To achieve better cross-lingual transfer performance, we propose the cross-lingual manifold mixup (X-Mixup) method, which adaptively calibrates the representation discrepancy and gives a compromised representation for target languages. Experiments on the XTREME benchmark show X-Mixup achieves 1.8% performance gains on multiple text understanding tasks, compared with strong baselines, and significantly reduces the cross-lingual representation discrepancy.
****

Evolutionary Diversity Optimization with Clustering-based Selection for Reinforcement Learning

Yutong Wang,Ke Xue,Chao Qian

Reinforcement Learning (RL) has achieved significant successes, which aims to obtain a single policy maximizing the expected cumulative rewards for a given task. However, in many real-world scenarios, e.g., navigating in complex environments and controlling robots, one may need to find a set of policies having both high rewards and diverse behaviors, which can bring better exploration and robust few-shot adaptation. Recently, some methods have been developed by using evolutionary techniques, including iterative reproduction and selection of policies. However, due to the inefficient selection mechanisms, these methods cannot fully guarantee both high quality and diversity. In this paper, we propose EDO-CS, a new Evolutionary Diversity Optimization algorithm with Clustering-based Selection. In each iteration, the policies are divided into several clusters based on their behaviors, and a high-quality policy is selected from each cluster for reproduction. EDO-CS also adaptively balances the importance between quality and diversity in the reproduction process. Experiments on various (i.e., deceptive and multi-modal) continuous control tasks, show the superior performance of EDO-CS over previous methods, i.e., EDO-CS can achieve a set of policies with both high quality and diversity efficiently while previous methods cannot.

****

CURVATURE-GUIDED DYNAMIC SCALE NETWORKS FOR MULTI-VIEW STEREO

Khang Truong Giang,Soohwan Song,Sungho Jo

Multi-view stereo (MVS) is a crucial task for precise 3D reconstruction. Most recent studies tried to improve the performance of matching cost volume in MVS by introducing a skilled design to cost formulation or cost regularization. In this paper, we focus on learning robust feature extraction to enhance the performance of matching costs, without need of heavy computation in the other steps. In particular, we present a dynamic scale feature extraction network, namely, CDSFNet. It is composed of multiple novel convolution layers, each of which can select a proper patch scale for each pixel guided by the normal curvature of image surface. As a result, CDFSNet can estimate the optimal patch scales to learn discriminative features for accurate matching computation between reference and source images. By combining the robust extracted features with an appropriate cost formulation strategy, our final MVS architecture can estimate depth maps more precisely. Extensive experiments showed that the proposed method outperforms other state-of-the-art methods on complex outdoor scenes. It significantly improves the completeness of reconstructed models. Moreover, the method can process the high resolution with faster run-time and lower memory compared to the other MVS methods.

****

Near-optimal Offline Reinforcement Learning with Linear Representation: Leveraging Variance Information with Pessimism

Ming Yin,Yaqi Duan,Mengdi Wang,Yu-Xiang Wang

Offline reinforcement learning, which seeks to utilize offline/historical data to optimize sequential decision-making strategies, has gained surging prominence in recent studies. Due to the advantage that appropriate function approximators can help mitigate the sample complexity burden in modern reinforcement learning problems, existing endeavors usually enforce powerful function representation models (e.g. neural networks) to learn the optimal policies. However, a precise understanding of the statistical limits with function representations, remains elusive, even when such a representation is linear.

Towards this goal, we study the statistical limits of offline reinforcement learning with linear model representations. To derive the tight offline learning bound, we design the variance-aware pessimistic value iteration (VAPVI), which adopts the conditional variance information of the value function for time-inhomogeneous episodic linear Markov decision processes (MDPs). VAPVI leverages estimated variances of the value functions to reweight the Bellman residuals in the least-square pessimistic value iteration and provides improved offline learning bounds over the best-known existing results (whereas the Bellman residuals are equally weighted by design). More importantly, our learning bounds are expressed in terms of system quantities, which provide natural instance-dependent characterizations that previous results are short of. We hope our results draw a clearer picture of what offline learning should look like when linear representations are provided.

****

Weakly-Supervised Learning of Disentangled and Interpretable Skills for Hierarchical Reinforcement Learning

Wonil Song,Sangryul Jeon,Hyesong Choi,Kwanghoon Sohn,Dongbo Min

Hierarchical reinforcement learning (RL) usually requires task-agnostic and interpretable skills that can be applicable to various downstream tasks. While many recent works have been proposed to learn such skills for a policy in unsupervised manner, the learned skills are still uninterpretable. To alleviate this, we propose a novel WEakly-supervised learning approach for learning Disentangled and Interpretable Skills (WEDIS) from the continuous latent representations of traje ctories. We accomplish this by extending a trajectory variational autoencoder (VAE) to impose an inductive bias with weak labels, which explicitly enforces the trajectory representations to be disentangled into factors of interest that we intend the model to learn. Given the latent representations as skills, a skill-based policy network is trained to generate similar trajectories to the learned decoder of the trajectory VAE. Additionally, we propose to train a policy network with single-step transitions and perform the trajectory-level behaviors at test time with the knowledge on the skills, which simplifies the exploration problem in the training. With a sample-efficient planning strategy based on the skills, we demonstrate that our method is effective in solving the hierarchical RL problems in experiments on several challenging navigation tasks with a long horizon and sparse rewards.

****

Exploring extreme parameter compression for pre-trained language models

Benyou Wang,Yuxin Ren,Lifeng Shang,Xin Jiang,Qun Liu

Recent work explored the potential of large-scale Transformer-based pre-trained models, especially Pre-trained Language Models (PLMs) in natural language processing. This raises many concerns from various perspectives, e.g., financial costs and carbon emissions.

Compressing PLMs like BERT with negligible performance loss for faster inference and cheaper deployment has attracted much attention. In this work, we aim to explore larger compression ratios for PLMs, among which tensor decomposition is a potential but under-investigated one. By comparing existing decomposition methods, Tucker decomposition is found to be parameter-efficient for compression. Two decomposition and reconstruction protocols are further proposed to improve the effectiveness and efficiency of Tucker decomposition in parameter compression.

Our compressed BERT with ${1}/{7}$ parameters in Transformer layers performs on-par with, sometimes slightly better than the original BERT in GLUE benchmark. A tiny version achieves 96.7\% performance of BERT-base with $ {1}/{48} $ encoder parameters (i.e., less than 2M parameters excluding the embedding layer) and \textbf{$2.7 \times$} faster on inference. To show that the proposed method is orthogonal to existing compression methods like knowledge distillation, we also explore the benefit of the proposed method on a distilled BERT.

****

Local Feature Swapping for Generalization in Reinforcement Learning

David Bertoin,Emmanuel Rachelson

Over the past few years, the acceleration of computing resources and research in Deep Learning has led to significant practical successes in a range of tasks, including in particular in computer vision. Building on these advances, reinforcement learning has also seen a leap forward with the emergence of agents capable of making decisions directly from visual observations. Despite these successes, the over-parametrization of neural architectures leads to memorization of the data used during training and thus to a lack of generalization.

Reinforcement learning agents based on visual inputs also suffer from this phenomenon by erroneously correlating rewards with unrelated visual features such as background elements. To alleviate this problem, we introduce a new regularization layer consisting of channel-consistent local permutations (CLOP) of the feature maps. The proposed permutations induce robustness to spatial correlations and help prevent overfitting behaviors in RL. We demonstrate, on the OpenAI Procgen Benchmark, that RL agents trained with the CLOP layer exhibit robustness to visual changes and better generalization properties than agents trained using other state-of-the-art regularization techniques.

****

Accelerating HEP simulations with Neural Importance Sampling

Nicolas Deutschmann,Niklas Götz

Virtually all high-energy-physics (HEP) simulations for the LHC rely on Monte Carlo using importance sampling by means of the VEGAS algorithm. However, complex high-precision calculations have become a challenge for the standard toolbox.

As a result, there has been keen interest in HEP for modern machine learning to power adaptive sampling. Despite previous work proving that normalizing-flow-pow ered neural importance sampling (NIS) sometimes outperforms VEGAS, existing research has still left major questions open, which we intend to solve by introducing ZüNIS, a fully automated NIS library.

We first show how to extend the original formulation of NIS to reuse samples over multiple gradient steps, yielding a significant improvement for slow functions. We then benchmark ZüNIS over a range of problems and show high performance with limited fine-tuning. This is crucial for ZüNIS to be a mature tool for the wider HEP public. We outline how the the library allows for non-experts to employ it with minimal effort, an essential condition to widely assess the value of NIS for LHC simulations.

****

Open-vocabulary Object Detection via Vision and Language Knowledge Distillation

Xiuye Gu,Tsung-Yi Lin,Weicheng Kuo,Yin Cui

We aim at advancing open-vocabulary object detection, which detects objects described by arbitrary text inputs. The fundamental challenge is the availability of training data. It is costly to further scale up the number of classes contained in existing object detection datasets. To overcome this challenge, we propose ViLD, a training method via Vision and Language knowledge Distillation. Our method distills the knowledge from a pretrained open-vocabulary image classification model (teacher) into a two-stage detector (student). Specifically, we use the teacher model to encode category texts and image regions of object proposals. Then we train a student detector, whose region embeddings of detected boxes are aligned with the text and image embeddings inferred by the teacher. We benchmark on LVIS by holding out all rare categories as novel categories that are not seen during training. ViLD obtains 16.1 mask APr with a ResNet-50 backbone, even outperforming the supervised counterpart by 3.8. When trained with a stronger teacher model ALIGN, ViLD achieves 26.3 APr. The model can directly transfer to other datasets without finetuning, achieving 72.2 AP50 on PASCAL VOC, 36.6 AP on COCO and 11.8 AP on Objects365. On COCO, ViLD outperforms the previous state-of-the-art (Zareian et al., 2021) by 4.8 on novel AP and 11.4 on overall AP. Code and demo are open-sourced at https://github.com/tensorflow/tpu/tree/master/models/official/detection/projects/vild.

****

Reinforcement Learning with Ex-Post Max-Min Fairness

Wang Chi Cheung,Zi Yi Ewe

We consider reinforcement learning with vectorial rewards, where the agent receives a vector of $K\geq 2$ different types of rewards at each time step. The agent aims to maximize the minimum total reward among the K reward types. Different from existing works that focus on maximizing the minimum expected total reward, i.e. \emph{ex-ante max-min fairness}, we maximize the expected minimum total reward, i.e. \emph{ex-post max-min fairness}. Through an example and numerical experiments, we show that the optimal policy for the former objective generally does not converge to optimality under the latter, even as the number of time steps T grows. Our main contribution is a novel algorithm, Online-ReOpt, that achieves near-optimality under our objective, assuming an optimization oracle that returns a near-optimal policy given any scalar reward. The expected objective value under Online-ReOpt is shown to converge to the asymptotic optimum as T increases. Finally, we propose offline variants to ease the burden of online computation in Online-ReOpt, and we propose generalizations from the max-min objective to concave utility maximization.

****

CLOOB: Modern Hopfield Networks with InfoLOOB Outperform CLIP

Andreas Fürst,Elisabeth Rumetshofer,Viet Thuong Tran,Hubert Ramsauer,Fei Tang,Johannes Lehner,D P Kreil,Michael K Kopp,Günter Klambauer,Angela Bitto-Nemling,Sepp Hochreiter

Contrastive learning with the InfoNCE objective is exceptionally successful in various self-supervised learning tasks. Recently, the CLIP model yielded impressive results on zero-shot transfer learning when using InfoNCE for learning visual representations from natural language supervision. However, InfoNCE as a lower bound on the mutual information has been shown to perform poorly for high mutual information. In contrast, the InfoLOOB upper bound (leave one out bound) works well for high mutual information but suffers from large variance and instabilities. We introduce "Contrastive Leave One Out Boost" (CLOOB), where modern Hopfield networks boost learning with the InfoLOOB objective. Modern Hopfield networks replace the original embeddings by retrieved embeddings in the InfoLOOB objective. The retrieved embeddings give InfoLOOB two assets. Firstly, the retrieved embeddings stabilize InfoLOOB, since they are less noisy and more similar to one another than the original embeddings. Secondly, they are enriched by correlations, since the covariance structure of embeddings is reinforced through retrievals. We compare CLOOB to CLIP after learning 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.

****

Constituency Tree Representation for Argument Unit Recognition
Samuel Guilluy,Florian Méhats,Billal Chouli

The extraction of arguments from sentences is usually studied by considering only the neighbourhood dependencies of words. Such a representation does not rely on the syntactic structure of the sentence and can lead to poor results especially in languages where grammatical categories are scattered in the sentence. In this paper, we investigate the advantages of using a constituency tree representation of sentences for argument discourse unit (ADU) prediction. We demonstrate that the constituency structure is more powerful than simple linear dependencies between neighbouring words in the sentence. Our work was organised as follows: First, we compare the maximum depth allowed for our constituency trees. This first step allows us to choose an optimal maximum depth. Secondly, we combine this structure with graph neural networks, which are very successful in neural network tasks. Finally, we evaluate the benefits of adding a conditional random field to model global dependencies between labels, given local dependency rules. We improve the current best models for argument unit recognition at token level and also present an explainability method to evaluate the suitability of our model architecture.

****

BioLCNet: Reward-modulated Locally Connected Spiking Neural Networks
Hafez Ghaemi,Erfan Mirzaei,Mahbod Nouri,Saeed Reza Kheradpisheh

Recent studies have shown that convolutional neural networks (CNNs) are not the only feasible solution for image classification. Furthermore, weight sharing and backpropagation used in CNNs do not correspond to the mechanisms present in the primate visual system. To propose a more biologically plausible solution, we designed a locally connected spiking neural network (SNN) trained using spike-timing-dependent plasticity (STDP) and its reward-modulated variant (R-STDP) learning rules. The use of spiking neurons and local connections along with reinforcement learning (RL) led us to the nomenclature BioLCNet for our proposed architecture. Our network consists of a rate-coded input layer followed by a locally connected hidden layer and a decoding output layer. A spike population-based voting scheme is adopted for decoding in the output layer. We used the MNIST dataset to obtain image classification accuracy and to assess the robustness of our rewarding system to varying target responses.

****

Confidence-aware Training of Smoothed Classifiers for Certified Robustness
Jongheon Jeong,Seojin Kim,Jinwoo Shin

Any classifier can be "smoothed out" under Gaussian noise to build a new classifier that is provably robust to ℓ_2-adversarial perturbations, viz., by averaging its predictions over the noise, namely via randomized smoothing. Under the smoothed classifiers, the fundamental trade-off between accuracy and (adversarial) robustness has been well evidenced in the literature: i.e., increasing the robustness of a classifier for an input can be at the expense of decreased accuracy for some other inputs. In this paper, we propose a simple training method leveraging this trade-off for obtaining more robust smoothed classifiers, in particular, through a sample-wise control of robustness over the training samples. We enable this control feasible by investigating the correspondence between robustness and prediction confidence of smoothed classifiers: specifically, we propose to use the "accuracy under Gaussian noise" as an easy-to-compute proxy of adversarial robustness for each input. We differentiate the training objective depending on this proxy to filter out samples that are unlikely to benefit from the worst-case (adversarial) objective. Our experiments following the standard benchmarks consistently show that the proposed method, despite its simplicity, exhibits improved certified robustness upon existing state-of-the-art training methods.

****

Distributional Perturbation for Efficient Exploration in Distributional Reinforcement Learning

Tae Hyun Cho,Sungyeob Han,Heesoo Lee,Kyungjae Lee,Jungwoo Lee

Distributional reinforcement learning aims to learn distribution of return under stochastic environments. Since the learned distribution of return contains rich information about the stochasticity of the environment, previous studies have relied on descriptive statistics, such as standard deviation, for optimism in face of uncertainty. These prior works are divided into risk-seeking or averse methods, which can be considered as having a one-sided tendency on risk. Unexpectedly, such approaches hinder convergence. In this paper, we propose a novel distributional reinforcement learning that explores by randomizing the risk criterion to reach a risk-neutral optimal policy. First, we provide a perturbed distributional Bellman optimality operator by distorting the risk measure in action selection. Second, we prove the convergence and optimality of the proposed method by using weaker contraction property. Our theoretical results support that the proposed method does not fall into biased exploration and converges to an optimal return distribution. Finally, we empirically show that our method outperforms other existing distribution-based algorithms in various environments including Atari games.

****

HydraSum - Disentangling Stylistic Features in Text Summarization using Multi-Decoder Models

Tanya Goyal,Nazneen Rajani,Wenhao Liu,Wojciech Maciej Kryscinski

Existing abstractive summarization models lack explicit control mechanisms that would allow users to influence the stylistic features of the model outputs. This results in generating generic summaries that do not cater to the users needs or preferences. To address this issue we introduce HydraSum, a new summarization architecture that extends the single decoder framework of current models, e.g. BART, to a mixture-of-experts version consisting of multiple decoders. Our proposed model encourages each expert, i.e. decoder, to learn and generate stylistically-distinct summaries along dimensions such as abstractiveness, length, specificity, and others. At each time step, HydraSum employs a gating mechanism that decides the contribution of each individual decoder to the next token's output probability distribution. Through experiments on three summarization datasets (CNN, Newsroom, XSum), we demonstrate that this gating mechanism automatically learns to assign contrasting summary styles to different HydraSum decoders under the standard training objective without the need for additional supervision. We further show that a guided version of the training process can explicitly govern which summary style is partitioned between decoders, e.g. high abstractiveness vs. low abstractiveness or high specificity vs. low specificity, and also increase the stylistic-difference between individual decoders. Finally, our experiments demonstrate that our decoder framework is highly flexible: during inference, we can sample from individual decoders or mixtures of different subsets of the decoders to yield a diverse set of summaries and enforce single- and multi-style control over summary generation.

****

Automatic Termination for Hyperparameter Optimization

Anastasia Makarova,Huibin Shen,Valerio Perrone,Aaron Klein,Jean Baptiste Faddoul,Andreas Krause,Matthias Seeger,Cedric Archambeau

Bayesian optimization (BO) is a widely popular approach for the hyperparameter optimization (HPO) of machine learning algorithms. At its core, BO iteratively ev aluates promising configurations until a user-defined budget, such as wall-clock time or number of iterations, is exhausted. While the final performance after tuning heavily depends on the provided budget, it is hard to pre-specify an optimal value in advance. In this work, we propose an effective and intuitive termination criterion for BO that automatically stops the procedure if it is sufficiently close to the global optima. Across an extensive range of real-world HPO problems, we show that our termination criterion achieves better test performance compared to existing baselines from the literature, such as stopping when the probability of improvement drops below a fixed threshold. We also provide evidence that these baselines are, compared to our method, highly sensitive to the choices of their own hyperparameters. Additionally, we find that overfitting might occur in the context of HPO, which is arguably an overlooked problem in the literature, and show that our termination criterion mitigates this phenomenon on both small and large datasets.

****

Domain-wise Adversarial Training for Out-of-Distribution Generalization

Shiji Xin,Yifei Wang,Jingtong Su,Yisen Wang

Despite the impressive success on many tasks, deep learning models are shown to rely on spurious features, which will catastrophically fail when generalized to out-of-distribution (OOD) data. To alleviate this issue, Invariant Risk Minimization (IRM) is proposed to extract domain-invariant features for OOD generalization. Nevertheless, recent work shows that IRM is only effective for a certain type of distribution shift (e.g., correlation shift) while fails for other cases (e.g., diversity shift). Meanwhile, another thread of method, Adversarial Training (AT), has shown better domain transfer performance, suggesting that it is potential to be an effective candidate for extracting domain-invariant features. In this paper, we investigate this possibility by exploring the similarity between the IRM and AT objectives. Inspired by this connection, we propose Domain-wise Adversarial Training (DAT), an AT-inspired method for alleviating distribution shift by domain-specific perturbations. Extensive experiments show that our proposed DAT can effectively remove the domain-varying features and improve OOD generalization on both correlation shift and diversity shift tasks.

****

Graph Kernel Neural Networks

Luca Cosmo,Giorgia Minello,Michael M. Bronstein,Emanuele Rodolà,Luca Rossi,Andrea Torsello

The convolution operator at the core of many modern neural architectures can effectively be seen as performing a dot product between an input matrix and a filter. While this is readily applicable to data such as images, which can be represented as regular grids in the Euclidean space, extending the convolution operator to work on graphs proves more challenging, due to their irregular structure. In this paper, we propose to use graph kernels, i.e., kernel functions that compute an inner product on graphs, to extend the standard convolution operator to the graph domain. This allows us to define an entirely structural model that does not require computing the embedding of the input graph. Our architecture allows to plug-in any type and number of graph kernels and has the added benefit of providing some interpretability in terms of the structural masks that are learned during the training process, similarly to what happens for convolutional masks in traditional convolutional neural networks. We perform an extensive ablation study to investigate the impact of the model hyper-parameters and we show that our model achieves competitive performance on standard graph classification datasets.

****

Particle Based Stochastic Policy Optimization

Qiwei Ye,Yuxuan Song,Chang Liu,Fangyun Wei,Tao Qin,Tie-Yan Liu

Stochastic polic have been widely applied for their good property in exploration and uncertainty quantification. Modeling policy distribution by joint state-action distribution within the exponential family has enabled flexibility in exploration and learning multi-modal policies and also involved the probabilistic perspective of deep reinforcement learning (RL). The connection between probabilistic inference and RL makes it possible to leverage the advancements of probabilis tic optimization tools. However, recent efforts are limited to the minimization of reverse KLdivergence which is confidence-seeking and may fade the merit of a stochastic policy. To leverage the full potential of stochastic policy and provide more flexible property, there is a strong motivation to consider different update rules during policy optimization. In this paper, we propose a particle-based probabilistic pol-icy optimization framework, ParPI, which enables the usage of a broad family of divergence or distances, such asf-divergences, and the Wasserstein distance which could serve better probabilistic behavior of the learned stochastic policy. Experiments in both online and offline settings demonstrate the effectiveness of the proposed algorithm as well as the characteristics of different discrepancy measures for policy optimization.

****

TRGP: Trust Region Gradient Projection for Continual Learning

Sen Lin,Li Yang,Deliang Fan,Junshan Zhang

Catastrophic forgetting is one of the major challenges in continual learning. To address this issue, some existing methods put restrictive constraints on the optimization space of the new task for minimizing the interference to old tasks. However, this may lead to unsatisfactory performance for the new task, especially when the new task is strongly correlated with old tasks. To tackle this challenge, we propose Trust Region Gradient Projection (TRGP) for continual learning to facilitate the forward knowledge transfer based on an efficient characterization of task correlation. Particularly, we introduce a notion of 'trust region' to select the most related old tasks for the new task in a layer-wise and single-shot manner, using the norm of gradient projection onto the subspace spanned by task inputs. Then, a scaled weight projection is proposed to cleverly reuse the frozen weights of the selected old tasks in the trust region through a layer-wise scaling matrix. By jointly optimizing the scaling matrices and the model, where the model is updated along the directions orthogonal to the subspaces of old tasks, TRGP can effectively prompt knowledge transfer without forgetting. Extensive experiments show that our approach achieves significant improvement over related state-of-the-art methods.

****

Learning from One and Only One Shot

Haizi Yu,Igor Mineyev,Lav R. Varshney,James Evans

Humans can generalize from one or a few examples, and even from very little pre-training on similar tasks. Machine learning (ML) algorithms, however, typically require large data to either learn or pre-learn to transfer. Inspired by nativism, we directly model very basic human innate priors in abstract visual tasks like character or doodle recognition. The result is a white-box model that learns transformation-based topological similarity akin to how a human would naturally and unconsciously ``distort'' an object when first seeing it. Using the simple Nearest-Neighbor classifier in this similarity space, our model approaches human-level character recognition using only one to ten examples per class and nothing else (no pre-training). This is in contrast to one-shot and few-shot settings that require significant pre-training. On standard benchmarks including MNIST, EMNIST-letters, and the harder Omniglot challenge, our model outperforms both neural-network-based and classical ML methods in the ``tiny-data'' regime, including few-shot learning models that use an extra background set to perform transfer learning. Moreover, mimicking simple clustering methods like k-means but in a non-Euclidean space, our model can adapt to an unsupervised setting and generate human-interpretable archetypes of a class.

****

Debiasing Pretrained Text Encoders by Paying Attention to Paying Attention

Yacine GACI,Boualem Benatallah,Fabio Casati,Khalid Benabdeslem

Recent studies in fair Representation Learning have observed a strong inclination for natural language processing (NLP) models to exhibit discriminatory stereotypes across gender, religion, race and many such social constructs. In comparison to the progress made in reducing bias from static word embeddings, fairness in sentence-level text encoders received little consideration despite their wider applicability in contemporary NLP tasks. In this paper, we propose a debiasing m ethod for pre-trained text encoders that both reduces social stereotypes, and inflicts next to no semantic offset. Unlike previous studies that directly manipulate the embeddings, we suggest to dive deeper into the operation of these encoders, and pay more attention to the way they pay attention to different social groups. We find that the attention mechanism is the root of all stereotypes. Then, we work on model debiasing by redistributing the attention scores of a text encoder such that it forgets any preference to historically advantaged groups, and attends to all social classes with the same intensity. Our experiments confirm that we successfully reduce bias with little damage to semantic representation.

****

Model-Based Offline Meta-Reinforcement Learning with Regularization

Sen Lin,Jialin Wan,Tengyu Xu,Yingbin Liang,Junshan Zhang

Existing offline reinforcement learning (RL) methods face a few major challenges, particularly the distributional shift between the learned policy and the behavior policy. Offline Meta-RL is emerging as a promising approach to address these challenges, aiming to learn an informative meta-policy from a collection of tasks. Nevertheless, as shown in our empirical studies, offline Meta-RL could be outperformed by offline single-task RL methods on tasks with good quality of datasets, indicating that a right balance has to be delicately calibrated between "exploring" the out-of-distribution state-actions by following the meta-policy and "exploiting" the offline dataset by staying close to the behavior policy. Motivated by such empirical analysis, we propose model-based offline $\text{\bf Me}$ta-RL with $\text{\bf r}$egularized $\text{\bf P}$olicy $\text{\bf O}$ptimization (MerPO), which learns a meta-model for efficient task structure inference and an informative meta-policy for safe exploration of out-of-distribution state-actions. In particular, we devise a new meta-Regularized model-based Actor-Critic (RAC) method for within-task policy optimization, as a key building block of MerPO, using both conservative policy evaluation and regularized policy improvement; and the intrinsic tradeoff therein is achieved via striking the right balance between two regularizers, one based on the behavior policy and the other on the meta-policy. We theoretically show that the learnt policy offers guaranteed improvement over both the behavior policy and the meta-policy, thus ensuring the performance improvement on new tasks via offline Meta-RL. Our experiments corroborate the superior performance of MerPO over existing offline Meta-RL methods.

****

Scale Mixtures of Neural Network Gaussian Processes

Hyungi Lee,Eunggu Yun,Hongseok Yang,Juho Lee

Recent works have revealed that infinitely-wide feed-forward or recurrent neural networks of any architecture correspond to Gaussian processes referred to as NNGP. While these works have extended the class of neural networks converging to Gaussian processes significantly, however, there has been little focus on broadening the class of stochastic processes that such neural networks converge to. In this work, inspired by the scale mixture of Gaussian random variables, we propose the scale mixture of NNGP for which we introduce a prior distribution on the scale of the last-layer parameters. We show that simply introducing a scale prior on the last-layer parameters can turn infinitely-wide neural networks of any architecture into a richer class of stochastic processes. With certain scale priors, we obtain heavy-tailed stochastic processes, and in the case of inverse gamma priors, we recover Student’s t processes. We further analyze the distributions of the neural networks initialized with our prior setting and trained with gradient descents and obtain similar results as for NNGP. We present a practical posterior-inference algorithm for the scale mixture of NNGP and empirically demonstrate its usefulness on regression and classification tasks. In particular, we show that in both tasks, the heavy-tailed stochastic processes obtained from our framework are robust to out-of-distribution data.

****

An Efficient and Reliable Tolerance-Based Algorithm for Principal Component Analysis

Michael Yeh,Ming Gu

Principal component analysis (PCA) is an important method for dimensionality red uction in data science and machine learning. But, it is expensive for large matrices when only a few principal components are needed. Existing fast PCA algorithms typically assume the user will supply the number of components needed, but in practice, they may not know this number beforehand. Thus, it is important to have fast PCA algorithms depending on a tolerance. For $m\times n$ matrices where a few principal components explain most of the variance in the data, we develop one such algorithm that runs in $O(mnl)$ time, where $l\ll \min(m,n)$ is a small multiple of the number of principal components. We provide approximation error bounds that are within a constant factor away from optimal and demonstrate its utility with data from a variety of applications.

****

A Johnson-Lindenstrauss Framework for Randomly Initialized CNNs

Ido Nachum,Jan Hazla,Michael Gastpar,Anatoly Khina

How does the geometric representation of a dataset change after the application of each randomly initialized layer of a neural network? The celebrated Johnson-Lindenstrauss lemma answers this question for linear fully-connected neural networks (FNNs), stating that the geometry is essentially preserved. For FNNs with the ReLU activation, the angle between two input contracts according to a known mapping. The question for non-linear convolutional neural networks (CNNs) becomes much more intricate. To answer this question, we introduce a geometric framework. For linear CNNs, we show that the Johnson--Lindenstrauss lemma continues to hold, namely, that the angle between two inputs is preserved. For CNNs with ReLU activation, on the other hand, the behavior is richer: The angle between the outputs contracts, where the level of contraction depends on the nature of the inputs. In particular, after one layer, the geometry of natural images is essentially preserved, whereas for Gaussian correlated inputs, CNNs exhibit the same contracting behavior as FNNs with ReLU activation.

****

Hindsight: Posterior-guided training of retrievers for improved open-ended generation

Ashwin Paranjape,Omar Khattab,Christopher Potts,Matei Zaharia,Christopher D Manning

Many text generation systems benefit from retrieving passages from a textual knowledge corpus (e.g., Wikipedia) and using them to generate the output. For open-ended generation tasks, like generating informative utterances in conversations, many varied passages z are relevant to the context x but few are relevant to the observed next utterance y (label). For such tasks, existing methods (that jointly train the retriever and generator) underperform: during training the top-k context-relevant retrieved passages might not contain the label-relevant passage and the generator may hence not learn a preference to ground its generated output in them. We propose using an additional guide-retriever that also conditions on the observed label y and “in hindsight” retrieves label-relevant passages during training. We maximize the evidence lower bound (ELBo) to jointly train the guide-retriever $Q(z|x,y)$ with the standard retriever $P_\eta(z|x)$ and the generator $P_\theta(y|x,z)$ and find that ELBo has better inductive biases than prior work. For informative conversations from the Wizard of Wikipedia dataset, with our posterior-guided training, the retriever finds passages with higher relevance in the top-10 (23% relative improvement), the generator’s responses are more grounded in the retrieved passage (19% relative improvement) and the end-to-end system produces better overall output (6.4% relative improvement).

****

Self-Supervised Graph Neural Networks for Improved Electroencephalographic Seizure Analysis

Siyi Tang,Jared Dunnmon,Khaled Kamal Saab,Xuan Zhang,Qianying Huang,Florian Dubost,Daniel Rubin,Christopher Lee-Messer

Automated seizure detection and classification from electroencephalography (EEG) can greatly improve seizure diagnosis and treatment. However, several modeling challenges remain unaddressed in prior automated seizure detection and classification studies: (1) representing non-Euclidean data structure in EEGs, (2) accurately classifying rare seizure types, and (3) lacking a quantitative interpretabi

lity approach to measure model ability to localize seizures. In this study, we address these challenges by (1) representing the spatiotemporal dependencies in EEGs using a graph neural network (GNN) and proposing two EEG graph structures that capture the electrode geometry or dynamic brain connectivity, (2) proposing a self-supervised pre-training method that predicts preprocessed signals for the next time period to further improve model performance, particularly on rare seizure types, and (3) proposing a quantitative model interpretability approach to assess a model’s ability to localize seizures within EEGs. When evaluating our approach on seizure detection and classification on a large public dataset (5,499 EEGs), we find that our GNN with self-supervised pre-training achieves 0.875 Area Under the Receiver Operating Characteristic Curve on seizure detection and 0.749 weighted F1-score on seizure classification, outperforming previous methods for both seizure detection and classification. Moreover, our self-supervised pre-training strategy significantly improves classification of rare seizure types (e.g. 47 points increase in combined tonic seizure accuracy over baselines). Furthermore, quantitative interpretability analysis shows that our GNN with self-supervised pre-training precisely localizes 25.4% focal seizures, a 21.9 point improvement over existing CNNs. Finally, by superimposing the identified seizure locations on both raw EEG signals and EEG graphs, our approach could provide clinicians with an intuitive visualization of localized seizure regions.
****
Self-supervised Learning for Sequential Recommendation with Model Augmentation
Zhiwei Liu,Yongjun Chen,Jia Li,Man Luo,Philip S. Yu,Caiming Xiong

The sequential recommendation aims at predicting the next items in user behaviors, which can be solved by characterizing item relationships in sequences. Due to the data sparsity and noise issues in sequences, a new self-supervised learning (SSL) paradigm is proposed to improve the performance, which employs contrastive learning between positive and negative views of sequences.

However, existing methods all construct views by adopting augmentation from data perspectives, while we argue that 1) optimal data augmentation methods are hard to devise, 2) data augmentation methods destroy sequential correlations, and 3) data augmentation fails to incorporate comprehensive self-supervised signals.

Therefore, we investigate the possibility of model augmentation to construct view pairs. We propose three levels of model augmentation methods: neuron masking, layer dropping, and encoder complementing.

This work opens up a novel direction in constructing views for contrastive SSL.

Experiments verify the efficacy of model augmentation for the SSL in the sequential recommendation.

****
Self-supervised regression learning using domain knowledge: Applications to improving self-supervised image denoising
Il Yong Chun,Dongwon Park,Xuehang Zheng,Se Young Chun,Yong Long

Regression that predicts continuous quantity is a central part of applications using computational imaging and computer vision technologies. Yet, studying and understanding self-supervised learning for regression tasks -- except for a particular regression task, image denoising -- have lagged behind. This paper proposes a general self-supervised regression learning (SSRL) framework that enables learning regression neural networks with only input data (but without ground-truth target data), by using a designable operator that encapsulates domain knowledge of a specific application. The paper underlines the importance of domain knowledge by showing that under some mild conditions, the better designable operator is used, the proposed SSRL loss becomes closer to ordinary supervised learning loss. Numerical experiments for natural image denoising and low-dose computational tomography denoising demonstrate that proposed SSRL significantly improves the denoising quality over several existing self-supervised denoising methods.
****
Group-based Interleaved Pipeline Parallelism for Large-scale DNN Training
PengCheng Yang,Xiaoming Zhang,Wenpeng Zhang,Ming Yang,Hong Wei

The recent trend of using large-scale deep neural networks (DNN) to boost perfor mance has propelled the development of the parallel pipelining technique for efficient DNN training, which has resulted in the development of several prominent pipelines such as GPipe, PipeDream, and PipeDream-2BW. However, the current leading pipeline PipeDream-2BW still suffers from two major drawbacks, i.e., the excessive memory redundancy and the delayed weight updates across all stages. In this work, we propose a novel pipeline named WPipe, which achieves better memory efficiency and fresher weight updates. WPipe uses a novel pipelining scheme that divides model partitions into two groups. It moves the forward pass of the next period of weight updates to the front of the backward pass of the current period of weight updates in the first group, retains the order in the second group, and updates each group alternatively. This scheme can eliminate half of the delayed gradients and memory redundancy compared to PipeDream-2BW. The experiments, which train large BERT language models, show that compared to PipeDream-2BW, WPipe achieves $1.4\times$ acceleration and reduces the memory footprint by 36%, without nearly sacrificing any final model accuracy.

****

Independent Component Alignment for Multi-task Learning

Dmitry Senushkin,Iaroslav Melekhov,Mikhail Romanov,Anton Konushin,Juho Kannala,Arno Solin

We present a novel gradient-based multi-task learning (MTL) approach that balances training in multi-task systems by aligning the independent components of the training objective. In contrast to state-of-the-art MTL approaches, our method is stable and preserves the ratio of highly correlated tasks gradients. The method is scalable, reduces overfitting, and can seamlessly handle multi-task objectives with a large difference in gradient magnitudes. We demonstrate the effectiveness of the proposed approach on a variety of MTL problems including digit classification, multi-label image classification, camera relocalization, and scene understanding. Our approach performs favourably compared to other gradient-based adaptive balancing methods, and its performance is backed up by theoretical analysis.

****

Understanding Square Loss in Training Overparametrized Neural Network Classifiers

Tianyang Hu,Jun Wang,Wenjia Wang,Zhenguo Li

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.

****

Minimax Optimality (Probably) Doesn't Imply Distribution Learning for GANs

Sitan Chen,Jerry Li,Yuanzhi Li,Raghu Meka

Arguably the most fundamental question in the theory of generative adversarial networks (GANs) is to understand when GANs can actually learn the underlying dist ribution. Theoretical and empirical evidence (see e.g. Arora-Risteski-Zhang '18) suggest local optimality of the empirical training objective is insufficient, yet it does not rule out the possibility that achieving a true population minimax optimal solution might imply distribution learning. In this paper, we show that standard cryptographic assumptions imply that this stronger condition is still insufficient. Namely, we show that if local pseudorandom generators (PRGs) exist, then for a large family of natural target distributions, there are ReLU network generators of constant depth and poly size which take Gaussian random seeds so that (i) the output is far in Wasserstein distance from the target distribution, but (ii) no polynomially large Lipschitz discriminator ReLU network can detect this. This implies that even achieving a population minimax optimal solution to the Wasserstein GAN objective is likely insufficient for distribution learning. Our techniques reveal a deep connection between GANs and PRGs, which we believe will lead to further insights into the computational landscape of GANs.

****

Decentralized Cross-Entropy Method for Model-Based Reinforcement Learning

Zichen Zhang,Jun Jin,Martin Jagersand,Jun Luo,Dale Schuurmans

Cross-Entropy Method (CEM) is a popular approach to planning in model-based reinforcement learning.

It has so far always taken a \textit{centralized} approach where the sampling distribution is updated \textit{centrally} based on the result of a top-k operation applied to \textit{all samples}.

We show that such a \textit{centralized} approach makes CEM vulnerable to local optima and impair its sample efficiency, even in a one-dimensional multi-modal optimization task.

In this paper, we propose \textbf{Decent}ralized \textbf{CEM (DecentCEM)} where an ensemble of CEM instances run independently from one another and each performs a local improvement of its own sampling distribution.

In the exemplar optimization task, the proposed decentralized approach DecentCEM finds the global optimum much more consistently than the existing CEM approaches that use either a single Gaussian distribution or a mixture of Gaussians.

Further, we extend the decentralized approach to sequential decision-making problems where we show in 13 continuous control benchmark environments that it matches or outperforms the state-of-the-art CEM algorithms in most cases, under the same budget of the total number of samples for planning.

****

Variational Inference for Discriminative Learning with Generative Modeling of Feature Incompletion

Kohei Miyaguchi,Takayuki Katsuki,Akira Koseki,Toshiya Iwamori

We are concerned with the problem of distributional prediction with incomplete features: The goal is to estimate the distribution of target variables given feature vectors with some of the elements missing. A typical approach to this problem is to perform missing-value imputation and regression, simultaneously or sequentially, which we call the generative approach. Another approach is to perform regression after appropriately encoding missing values into the feature, which we call the discriminative approach. In comparison, the generative approach is more robust to the feature corruption while the discriminative approach is more favorable to maximize the performance of prediction.

In this study, we propose a hybrid method to take the best of both worlds. Our method utilizes the black-box variational inference framework so that it can be applied to a wide variety of modern machine learning models, including the variational autoencoders. We also confirmed the effectiveness of the proposed method empirically.

****

Robust and Personalized Federated Learning with Spurious Features: an Adversarial Approach

Xiaoyang Wang,Han Zhao,Klara Nahrstedt,Oluwasanmi O Koyejo

A common approach for personalized federated learning is fine-tuning the global machine learning model to each local client. While this addresses some issues of statistical heterogeneity, we find that such personalization methods are often vulnerable to spurious features, leading to bias and diminished generalization performance. However, debiasing the personalized models under spurious features is difficult. To this end, we propose a strategy to mitigate the effect of spurious features based on our observation that the global model in the federated learning step has a low accuracy disparity due to statistical heterogeneity. Then, we estimate and mitigate the accuracy disparity of personalized models using the global model and adversarial transferability in the personalization step. Empirical results on MNIST, CelebA, and Coil20 datasets show that our method reduces the accuracy disparity of the personalized model on the bias-conflicting data samples from 15.12% to 2.15%, compared to existing personalization approaches, while preserving the benefit of enhanced average accuracy from fine-tuning.

****

Offline Reinforcement Learning with Value-based Episodic Memory

Xiaoteng Ma,Yiqin Yang,Hao Hu,Jun Yang,Chongjie Zhang,Qianchuan Zhao,Bin Liang,Qihan Liu

Offline reinforcement learning (RL) shows promise of applying RL to real-world problems by effectively utilizing previously collected data. Most existing offline RL algorithms use regularization or constraints to suppress extrapolation error for actions outside the dataset. In this paper, we adopt a different framework, which learns the V-function instead of the Q-function to naturally keep the learning procedure within the support of an offline dataset. To enable effective generalization while maintaining proper conservatism in offline learning, we propose Expectile V-Learning (EVL), which smoothly interpolates between the optimal value learning and behavior cloning. Further, we introduce implicit planning along offline trajectories to enhance learned V-values and accelerate convergence. Together, we present a new offline method called Value-based Episodic Memory (VEM). We provide theoretical analysis for the convergence properties of our proposed VEM method, and empirical results in the D4RL benchmark show that our method achieves superior performance in most tasks, particularly in sparse-reward tasks.

****

Momentum Conserving Lagrangian Neural Networks

Ravinder Bhattoo,Sayan Ranu,N M Anoop Krishnan

Realistic models of physical world rely on differentiable symmetries that, in turn, correspond to conservation laws. Recent works on Lagrangian and Hamiltonian neural networks show that the underlying symmetries of a system can be easily learned by a neural network when provided with an appropriate inductive bias. However, these models still suffer from issues such as inability to generalize to arbitrary system sizes, poor interpretability, and most importantly, inability to learn translational and rotational symmetries, which lead to the conservation laws of linear and angular momentum, respectively. Here, we present a momentum conserving Lagrangian neural network (MCLNN) that learns the Lagrangian of a system, while also preserving the translational and rotational symmetries. We test our approach on linear and non-linear spring systems, and a gravitational system, demonstrating the energy and momentum conservation. We also show that the model developed can generalize to systems of any arbitrary size. Finally, we discuss the interpretability of the MCLNN, which directly provides physical insights into the interactions of multi-particle systems.

****

MonoDistill: Learning Spatial Features for Monocular 3D Object Detection

Zhiyu Chong,Xinzhu Ma,Hong Zhang,Yuxin Yue,Haojie Li,Zhihui Wang,Wanli Ouyang

3D object detection is a fundamental and challenging task for 3D scene understanding, and the monocular-based methods can serve as an economical alternative to the stereo-based or LiDAR-based methods. However, accurately locating objects in the 3D space from a single image is extremely difficult due to the lack of spatial cues. To mitigate this issue, we propose a simple and effective scheme to introduce the spatial information from LiDAR signals to the monocular 3D detectors, without introducing any extra cost in the inference phase. In particular, we first project the LiDAR signals into the image plane and align them with the RGB

images. After that, we use the resulting data to train a 3D detector (LiDAR Net) using the same architecture as the baseline model. Finally, this LiDAR Net can serve as the teacher to transfer the learned knowledge to the baseline model. Experimental results show that the proposed method can significantly boost the performance of the baseline model and ranks the 1^{st} place among all monocular-based methods on the KITTI benchmark. Besides, extensive ablation studies are conducted, which further prove the effectiveness of each part of our designs and illustrate what the baseline model has learned from the LiDAR Net.
****

Object-Centric Neural Scene Rendering

Michelle Guo,Alireza Fathi,Jiajun Wu,Thomas Funkhouser

We present a method for composing photorealistic scenes from captured images of objects. Traditional computer graphics methods are unable to model objects from observations only; instead, they rely on underlying computer graphics models. Our work builds upon neural radiance fields (NeRFs), which implicitly model the volumetric density and directionally-emitted radiance of a scene. While NeRFs synthesize realistic pictures, they only model static scenes and are closely tied to specific imaging conditions. This property makes NeRFs hard to generalize to new scenarios, including new lighting or new arrangements of objects. Instead of learning a scene radiance field as a NeRF does, we propose to learn object-centric neural scattering functions (OSFs), a representation that models per-object light transport implicitly using a lighting- and view-dependent neural network. This enables rendering scenes even when objects or lights move, without retraining. Combined with a volumetric path tracing procedure, our framework is capable of rendering light transport effects including occlusions, specularities, shadows, and indirect illumination, both within individual objects and between different objects. We evaluate our approach on synthetic and real world datasets and generalize to novel scene configurations, producing photorealistic, physically accurate renderings of multi-object scenes.
****

Properties from mechanisms: an equivariance perspective on identifiable representation learning

Kartik Ahuja,Jason Hartford,Yoshua Bengio

A key goal of unsupervised representation learning is ``inverting'' a data generating process to recover its latent properties. Existing work that provably achieves this goal relies on strong assumptions on relationships between the latent variables (e.g., independence conditional on auxiliary information). In this paper, we take a very different perspective on the problem and ask, ``Can we instead identify latent properties by leveraging knowledge of the mechanisms that govern their evolution?'' We provide a complete characterization of the sources of non-identifiability as we vary knowledge about a set of possible mechanisms. In particular, we prove that if we know the exact mechanisms under which the latent properties evolve, then identification can be achieved up to any equivariances that are shared by the underlying mechanisms. We generalize this characterization to settings where we only know some hypothesis class over possible mechanisms, as well as settings where the mechanisms are stochastic. We demonstrate the power of this mechanism-based perspective by showing that we can leverage our results to generalize existing identifiable representation learning results. These results suggest that by exploiting inductive biases on mechanisms, it is possible to design a range of new identifiable representation learning approaches.
****

Online Tuning for Offline Decentralized Multi-Agent Reinforcement Learning

Jiechuan Jiang,Zongqing Lu

Offline reinforcement learning could learn effective policies from a fixed dataset, which is promising in real-world applications. However, in offline decentralized multi-agent reinforcement learning, due to the discrepancy between the behavior policy and learned policy, the transition dynamics in offline experiences do not accord with the transition dynamics in online execution, which creates severe errors in value estimates, leading to uncoordinated and suboptimal policies. One way to overcome the transition bias is to bridge offline training and onlin e tuning. However, considering both deployment efficiency and sample efficiency, we could only collect very limited online experiences, making it insufficient to use merely online data for updating the agent policy. To utilize both offline and online experiences to tune the policies of agents, we introduce online transition correction (OTC) to implicitly correct the biased transition dynamics by modifying sampling probabilities. We design two types of distances, i.e., embedding-based and value-based distance, to measure the similarity between transitions, and further propose an adaptive rank-based prioritization to sample transitions according to the transition similarity. OTC is simple yet effective to increase data efficiency and improve agent policies in online tuning. Empirically, we show that OTC outperforms baselines in a variety of tasks.

****

Revisiting Over-smoothing in BERT from the Perspective of Graph

Han Shi,JIAHUI GAO,Hang Xu,Xiaodan Liang,Zhenguo Li,Lingpeng Kong,Stephen M. S. Lee,James Kwok

Recently over-smoothing phenomenon of Transformer-based models is observed in both vision and language fields. However, no existing work has delved deeper to further investigate the main cause of this phenomenon. In this work, we make the attempt to analyze the over-smoothing problem from the perspective of graph, where such problem was first discovered and explored. Intuitively, the self-attention matrix can be seen as a normalized adjacent matrix of a corresponding graph. Based on the above connection, we provide some theoretical analysis and find that layer normalization plays a key role in the over-smoothing issue of Transformer-based models. Specifically, if the standard deviation of layer normalization is sufficiently large, the output of Transformer stacks will converge to a specific low-rank subspace and result in over-smoothing. To alleviate the over-smoothing problem, we consider hierarchical fusion strategies, which combine the representations from different layers adaptively to make the output more diverse. Extensive experiment results on various data sets illustrate the effect of our fusion method.

****

Offline Decentralized Multi-Agent Reinforcement Learning

Jiechuan Jiang,Zongqing Lu

In many real-world multi-agent cooperative tasks, due to high cost and risk, agents cannot continuously interact with the environment and collect experiences during learning, but have to learn from offline datasets. However, the transition probabilities calculated from the dataset can be much different from the transition probabilities induced by the learned policies of other agents, creating large errors in value estimates. Moreover, the experience distributions of agents' datasets may vary wildly due to diverse behavior policies, causing large difference in value estimates between agents. Consequently, agents will learn uncoordinated suboptimal policies. In this paper, we propose MABCQ, which exploits value deviation and transition normalization to modify the transition probabilities. Value deviation optimistically increases the transition probabilities of high-value next states, and transition normalization normalizes the biased transition probabilities of next states. They together encourage agents to discover potential optimal and coordinated policies. Mathematically, we prove the convergence of Q-learning under the non-stationary transition probabilities after modification. Empirically, we show that MABCQ greatly outperforms baselines and reduces the difference in value estimates between agents.

****

EXACT: Scalable Graph Neural Networks Training via Extreme Activation Compression

Zirui Liu,Kaixiong Zhou,Fan Yang,Li Li,Rui Chen,Xia Hu

Training Graph Neural Networks (GNNs) on large graphs is a fundamental challenge due to the high memory usage, which is mainly occupied by activations (e.g., node embeddings). Previous works usually focus on reducing the number of nodes retained in memory.

In parallel, unlike what has been developed for other types of neural networks, training with compressed activation maps is less explored for GNNs. This extensi on is notoriously difficult to implement due to the miss of necessary tools in common graph learning packages. To unleash the potential of this direction, we provide { an} optimized GPU implementation which supports training GNNs with compressed activations. Based on the implementation, we propose a memory-efficient framework called ``EXACT'', which for the first time demonstrate the potential and evaluate the feasibility of training GNNs with compressed activations. We systematically analyze the trade-off among the memory saving, time overhead, and accuracy drop. In practice, EXACT can reduce the memory footprint of activations by up to $32\times$ with 0.2-0.5% accuracy drop and 10-25% time overhead across different models and datasets. We implement EXACT as an extension for Pytorch Geometric and Pytorch. In practice, for Pytorch Geometric, EXACT can trim down the hardware requirement of training a three-layer full-batch GraphSAGE on \textit{ogbn-products} from a 48GB GPU to a 12GB GPU.

****

Self-Supervised Structured Representations for Deep Reinforcement Learning
Hyesong Choi,Hunsang Lee,Wonil Song,Sangryul Jeon,Kwanghoon Sohn,Dongbo Min

Recent reinforcement learning (RL) methods have found extracting high-level features from raw pixels with self-supervised learning to be effective in learning policies. However, these methods focus on learning global representations of images, and disregard local spatial structures present in the consecutively stacked frames. In this paper, we propose a novel approach that learns self-supervised structured representations ($\mathbf{S}^3$R) for effectively encoding such spatial structures in an unsupervised manner. Given the input frames, the structured latent volumes are first generated individually using an encoder, and they are used to capture the change in terms of spatial structures, i.e., flow maps among multiple frames. To be specific, the proposed method establishes flow vectors between two latent volumes via a supervision by the image reconstruction loss. This enables for providing plenty of local samples for training the encoder of deep RL. We further attempt to leverage the structured representations in the self-predictive representations (SPR) method that predicts future representations using the action-conditioned transition model. The proposed method imposes similarity constraints on the three latent volumes; warped query representations by estimated flows, predicted target representations from the transition model, and target representations of future state. Experimental results on complex tasks in Atari Games and DeepMind Control Suite demonstrate that the RL methods are significantly boosted by the proposed self-supervised learning of structured representations.

The code is available at https://sites.google.com/view/iclr2022-s3r.

****

Conditional set generation using Seq2seq models
Aman Madaan,Dheeraj Rajagopal,Antoine Bosselut,Yiming Yang

Conditional set generation learns a mapping from an input sequence of tokens to a set. Several popular natural language processing (NLP) tasks, such as entity typing and dialogue emotion tagging, are instances of set generation. Sequence-to-sequence models are a popular choice to model set generation but this typical approach of treating a set as a sequence does not fully leverage its key properties, namely order-invariance and cardinality. We propose a novel data augmentation approach that recovers informative orders for labels using their dependence information. Further, we jointly model the set cardinality and output by listing the set size as the first element and taking advantage of the autoregressive factorization used by seq2seq models. Our experiments in simulated settings and on three diverse NLP datasets show that our method improves over strong seq2seq baselines by about 9% on absolute F1 score. We will release all code and data upon acceptance.

****

Low-Precision Stochastic Gradient Langevin Dynamics
Ruqi Zhang,Andrew Gordon Wilson,Christopher De Sa

Low-precision optimization is widely used to accelerate large-scale deep learning. Despite providing better uncertainty estimation and generalization, sampling methods remain mostly unexplored in this space. In this paper, we provide the fi rst study of low-precision Stochastic Gradient Langevin Dynamics (SGLD), arguing that it is particularly suited to low-bit arithmetic due to its intrinsic ability to handle system noise. We prove the convergence of low-precision SGLD on strongly log-concave distributions, showing that with full-precision gradient accumulators, SGLD is more robust to quantization error than SGD; however, with low-precision gradient accumulators, SGLD can diverge arbitrarily far from the target distribution with small stepsizes. To remedy this issue, we develop a new quantization function that preserves the correct variance in each update step. We demonstrate that the resulting low-precision SGLD algorithm is comparable to full-precision SGLD and outperforms low-precision SGD on deep learning tasks.

****

Batch size-invariance for policy optimization

Jacob Hilton,Karl Cobbe,John Schulman

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.

****

A Branch and Bound Framework for Stronger Adversarial Attacks of ReLU Networks

Huan Zhang,Shiqi Wang,Kaidi Xu,Yihan Wang,Suman Jana,Cho-Jui Hsieh,J Zico Kolter

Strong adversarial attacks are important for evaluating the true robustness of deep neural networks. Most existing attacks find adversarial examples via searching in the input space, e.g., using gradient descent. In this work, we formulate an adversarial attack using a branch-and-bound (BaB) procedure on ReLU neural networks and search adversarial examples in the activation space corresponding to binary variables in a mixed integer programming (MIP) formulation. This attack formulation can be used to tackle hard instances where none of the existing adversarial attacks can succeed. Existing attacks using this formulation rely on generic solvers which cannot exploit the structure of neural networks and also cannot utilize GPU acceleration, so they are mostly limited to small networks and easy problem instances. To improve its scalability and practicability, we propose a top-down beam-search approach to quickly identify the subspace that may contain adversarial examples. The search utilizes the bound propagation based neural network verifiers on GPUs to rapidly evaluate a large number of searching regions, which is not possible in generic MIP solvers. Moreover, we exploit the fact that good candidates of adversarial examples can be easily found via gradient based attacks, and build an adversarial candidates pool to further guide the search in activation space via diving techniques. Additionally, any candidate adversarial examples found during the process are refined using a bottom-up large neighbourhood search (LNS) guided by the candidates pool. Our adversarial attack framework, BaB-Attack, opens up a new opportunity for designing novel adversarial attacks not limited to searching the input space, and enables us to borrow techniques from integer programming theory and neural network verification to build stronger attacks. In experiments, we can successfully generate adversarial examples for hard input instances where existing strong adversarial attacks fail, and outperform off-the-shelf MIP solver based attacks in both success rates and efficiency. Our results further close the gap between the upper bound of robust accuracy obtained by attacks and the lower bound obtained by verification.

****

Speech-MLP: a simple MLP architecture for speech processing

Chao Xing,Dong Wang,Lirong Dai,Qun Liu,Anderson Avila

Overparameterized transformer-based architectures have shown remarkable performance in recent years, achieving state-of-the-art results in speech processing tasks such as speech recognition, speech synthesis, keyword spotting, and speech en hancement et al. The main assumption is that with the underlying self-attention mechanism, transformers can ultimately capture the long-range temporal dependency from speech signals. In this paper, we propose a multi-layer perceptron (MLP) architecture, namely speech-MLP, useful for extracting information from speech signals. The model splits feature channels into non-overlapped chunks and processes each chunk individually. The processed chunks are then merged together and processed to consolidate the output. By setting the different numbers of chunks and focusing on different contextual window sizes, speech-MLP learns multiscale local temporal dependency. The proposed model is successfully evaluated on two tasks: keyword spotting and speech enhancement. In our experiments, we use two benchmark datasets for keyword spotting (Google speech command V2-35 and LibriWords) and the VoiceBank dataset for the speech enhancement task. In all experiments, speech-MLP surpassed transformer-based solutions, achieving state-of-the-art performance with fewer parameters and simpler training schemes. Such results indicate that oftentimes more complex models such as transformers are not necessary for speech processing tasks. Hence, they should not be considered as the first option as simpler and more compact models can offer optimal performance.

****

Provably convergent quasistatic dynamics for mean-field two-player zero-sum games

Chao Ma,Lexing Ying

In this paper, we study the problem of finding mixed Nash equilibrium for mean-field two-player zero-sum games. Solving this problem requires optimizing over two probability distributions. We consider a quasistatic Wasserstein gradient flow dynamics in which one probability distribution follows the Wasserstein gradient flow, while the other one is always at the equilibrium. Theoretical analysis are conducted on this dynamics, showing its convergence to the mixed Nash equilibrium under mild conditions. Inspired by the continuous dynamics of probability distributions, we derive a quasistatic Langevin gradient descent method with inner-outer iterations, and test the method on different problems, including training mixture of GANs.

****

W-CTC: a Connectionist Temporal Classification Loss with Wild Cards

Xingyu Cai,Jiahong Yuan,Yuchen Bian,Guangxu Xun,Jiaji Huang,Kenneth Church

Connectionist Temporal Classification (CTC) loss is commonly used in sequence learning applications. For example, in Automatic Speech Recognition (ASR) task, the training data consists of pairs of audio (input sequence) and text (output label),without temporal alignment information. Standard CTC computes a loss by aggregating over all possible alignment paths, that map the entire sequence to the entire label (full alignment). However, in practice, there are often cases where the label is incomplete. Specifically, we solve the partial alignment problem where the label only matches a middle part of the sequence. This paper proposes the wild-card CTC (W-CTC) to address this issue, by padding wild-cards at both ends of the labels. Consequently, the proposed W-CTC improves the standard CTC via aggregating over even more alignment paths. Evaluations on a number of tasks in speech and vision domains, show that the proposed W-CTC consistently outperforms the standard CTC by a large margin when label is incomplete. The effectiveness of the proposed method is further confirmed in an ablation study.

****

Bandit Learning with Joint Effect of Incentivized Sampling, Delayed Sampling Feedback, and Self-Reinforcing User Preferences

Tianchen Zhou,Jia Liu,Chaosheng Dong,Yi Sun

In this paper, we consider a new multi-armed bandit (MAB) framework motivated by three common complications in online recommender systems in practice: (i) the platform (learning agent) cannot sample an intended product directly and has to incentivize customers to select this product (e.g., promotions and coupons); (ii) customer feedbacks are often received later than their selection times; and (iii) customer preferences among products are influenced and reinforced by historical feedbacks. From the platform's perspective, the goal of the MAB framework is to maximize total reward without incurring excessive incentive costs. A major ch allenge of this MAB framework is that the loss of information caused by feedback delay complicates both user preference evolution and arm incentivizing decisions, both of which are already highly non-trivial even by themselves. Toward this end, we first propose a policy called ``UCB-Filtering-with-Delayed-Feedback'' (UCB-FDF) policy for this new MAB framework. In our analysis, we consider delayed feedbacks that can have either arm-independent or arm-dependent distributions. In both cases, we allow unbounded support for the random delays, i.e., the random delay can be infinite. We show that the delay impacts in both cases can still be upper bounded by an additive penalty on both the regret and total incentive costs. This further implies that logarithmic regret and incentive cost growth rates are achievable under this new MAB framework. Experimental results corroborate our theoretical analysis on both regret and incentive costs.

****

Guided-TTS:Text-to-Speech with Untranscribed Speech
Heeseung Kim,Sungwon Kim,Sungroh Yoon
Most neural text-to-speech (TTS) models require $\langle$speech, transcript$\rangle$ paired data from the desired speaker for high-quality speech synthesis, which limits the usage of large amounts of untranscribed data for training. In this work, we present Guided-TTS, a high-quality TTS model that learns to generate speech from untranscribed speech data. Guided-TTS combines an unconditional diffusion probabilistic model with a separately trained phoneme classifier for text-to-speech. By modeling the unconditional distribution for speech, our model can utilize the untranscribed data for training. For text-to-speech synthesis, we guide the generative process of the unconditional DDPM via phoneme classification to produce mel-spectrograms from the conditional distribution given transcript. We show that Guided-TTS achieves comparable performance with the existing methods without any transcript for LJSpeech. Our results further show that a single speaker-dependent phoneme classifier trained on multispeaker large-scale data can guide unconditional DDPMs for various speakers to perform TTS.
****

Evaluating Predictive Distributions: Does Bayesian Deep Learning Work?
Ian Osband,Zheng Wen,Seyed Mohammad Asghari,Xiuyuan Lu,Morteza Ibrahimi,Vikranth Dwaracherla,Dieterich Lawson,Brendan O'Donoghue,Botao Hao,Benjamin Van Roy
Posterior predictive distributions quantify uncertainties ignored by point estimates.
This paper introduces \textit{The Neural Testbed}, which provides tools for the systematic evaluation of agents that generate such predictions.
Crucially, these tools assess not only the quality of marginal predictions per input, but also joint predictions given many inputs.
Joint distributions are often critical for useful uncertainty quantification, but they have been largely overlooked by the Bayesian deep learning community.
We benchmark several approaches to uncertainty estimation using a neural-network-based data generating process.
Our results reveal the importance of evaluation beyond marginal predictions.
Further, they reconcile sources of confusion in the field, such as why Bayesian deep learning approaches that generate accurate marginal predictions perform poorly in sequential decision tasks, how incorporating priors can be helpful, and what roles epistemic versus aleatoric uncertainty play when evaluating performance.
We also present experiments on real-world challenge datasets, which show a high correlation with testbed results, and that the importance of evaluating joint predictive distributions carries over to real data.
As part of this effort, we opensource The Neural Testbed, including all implementations from this paper.
****

BLOOD: Bi-level Learning Framework for Out-of-distribution Generalization
Jun-Hyun Bae,Inchul Choi,Minho Lee
Empirical risk minimization (ERM) based machine learning algorithms have suffered from weak generalization performance on the out-of-distribution (OOD) data whe

n the training data are collected from separate environments with unknown spurious correlations. To address this problem, previous works either exploit prior human knowledge for biases in the dataset or apply the two-stage process, which re-weights spuriously correlated samples after they were identified by the biased classifier. However, most of them fail to remove multiple types of spurious correlations that exist in training data. In this paper, we propose a novel bi-level learning framework for OOD generalization, which can effectively remove multiple unknown types of biases without any prior bias information or separate re-training steps of a model. In our bi-level learning framework, we uncover spurious correlations in the inner-loop with shallow model-based predictions and dynamically re-group the data to leverage the group distributionally robust optimization method in the outer-loop, minimizing the worst-case risk across all batches. Our main idea applies the unknown bias discovering process to the group construction method of the group DRO algorithm in a bi-level optimization setting and provides a unified de-biasing framework that can handle multiple types of biases in data. In empirical evaluations on both synthetic and real-world datasets, our framework shows superior OOD performance compared to all other state-of-the-art OOD methods by a large margin. Furthermore, it successfully removes multiple types of biases in the training data groups that most other OOD models fail.

****

FoveaTer: Foveated Transformer for Image Classification

Aditya Jonnalagadda,William Yang Wang,B.S. Manjunath,Miguel Eckstein

Many animals and humans process the visual field with varying spatial resolution (foveated vision) and use peripheral processing to make eye movements and point the fovea to acquire high-resolution information about objects of interest. This architecture results in computationally efficient rapid scene exploration. Recent progress in vision Transformers has brought about new alternatives to the traditionally convolution-reliant computer vision systems. However, the Transformer models do not explicitly model the foveated properties of the visual system nor the interaction between eye movements and the classification task. We propose foveated Transformer (FoveaTer) model, which uses pooling regions and eye movements to perform object classification tasks using a vision Transformer architecture. Our proposed model pools the image features using squared pooling regions, an approximation to the biologically-inspired foveated architecture, and uses the pooled features as an input to a Transformer Network. It decides on subsequent fixation locations based on the attention assigned by the Transformer to various locations from previous and present fixations. The model uses a confidence threshold to stop scene exploration, dynamically allocating more fixation/computational resources to more challenging images. After reaching the stopping criterion, the model makes the final object category decision. We construct a Foveated model using our proposed approach and compare it against a Full-resolution model, which does not contain any pooling. On the ImageNet-100 dataset, our Foveated model achieves the accuracy of the Full-resolution model using only 35% transformer computations and 73% overall computations. Finally, we demonstrate our model's robustness against adversarial attacks, where it outperforms the full-resolution model.

****

Adversarial twin neural networks: maximizing physics recovery for physical system

Haoran Li,Erik Blasch,Jingyi Yuan,Yang Weng

The exact modeling of modern physical systems is challenging due to the expanding system territory and insufficient sensors. To tackle this problem, existing methods utilize sparse regression to find physical parameters or add another virtual learning model like a Neural Network (NN) to universally approximate the unobserved physical quantities. However, the two models can't perfectly play their own roles in joint learning without proper restrictions. Thus, we propose (1) sparsity regularization for the physical model and (2) physical superiority over the virtual model. They together define output boundaries for the physical and virtual models. Further, even the two models output properly, the joint model still can't guarantee learning maximal physical knowledge. For example, if the data o f an observed node can linearly represent those of an unobserved node, these two nodes can be aggregated. Therefore, we propose (3) to seek the dissimilarity of physical and virtual outputs to obtain maximal physics. To achieve goals (1)-(3), we design a twin structure of the Physical Neural Network (PNN) and Virtual Neural Network (VNN), where sparse regularization and skip-connections are utilized to guarantee (1) and (2). Then, we propose an adversarial learning scheme to maximize output dissimilarity, achieving (3). We denote the model as the Adversarial Twin Neural Network (ATN). Finally, we conduct extensive experiments over various systems to demonstrate the best performance of ATN over other state-of-the-art methods.

****

Learning Two-Step Hybrid Policy for Graph-Based Interpretable Reinforcement Learning

Tongzhou Mu,Kaixiang Lin,Feiyang Niu,Govind Thattai

We present a two-step hybrid reinforcement learning (RL) policy that is designed to generate interpretable and robust hierarchical policies
on the RL problem with graph-based input. Unlike prior deep reinforcement learning policies parameterized by an end-to-end black-box graph neural network, our approach disentangles the decision-making process into two steps. The first step is a simplified classification problem that maps the graph input to an action group where all actions share a similar semantic meaning. The second step implements a sophisticated rule-miner that conduct explicit one-hop reasoning over the graph and identifies decisive edges in the graph input without the necessity of heavy domain knowledge. This two-step hybrid policy presents human-friendly interpretations and achieves better performance in terms of generalization and robustness. Extensive experimental studies on four levels of complex text-based games have demonstrated the superiority of the proposed method compared to the state-of-the-art.

****

ZerO Initialization: Initializing Residual Networks with only Zeros and Ones

Jiawei Zhao,Florian Tobias Schaefer,Anima Anandkumar

Deep neural networks are usually initialized with random weights, with adequately selected initial variance to ensure stable signal propagation during training. However, there is no consensus on how to select the variance, and this becomes challenging especially as the number of layers grows. In this work, we replace the widely used random weight initialization with a fully deterministic initialization scheme ZerO, which initializes residual networks with only zeros and ones. By augmenting the standard ResNet architectures with a few extra skip connections and Hadamard transforms, ZerO allows us to start the training from zeros and ones entirely. This has many benefits such as improving reproducibility (by reducing the variance over different experimental runs) and allowing network training without batch normalization. Surprisingly, we find that ZerO achieves state-of-the-art performance over various image classification datasets, including ImageNet, which suggests random weights may be unnecessary for modern network initialization.

****

Learning to Model Editing Processes

Machel Reid,Graham Neubig

Most existing sequence generation models produce outputs in one pass, usually left-to-right. However, this is in contrast with a more natural approach that humans use in generating content; iterative refinement and editing. Recent work has introduced edit-based models for various tasks (such as neural machine translation and text style transfer), but these generally model a single edit. In this work, we propose to model editing processes, modeling the whole process of iteratively generating sequences. We form a conceptual framework to describe the likelihood of multi-step edits, and describe neural models that can learn a generative model of sequences based on these multi-step edits. We introduce baseline results and metrics on this task, finding that modeling editing processes improves performance on a variety of axes on both our proposed task and related downstream tasks compared to previous single-step models of edits.

****

Edge Partition Modulated Graph Convolutional Networks
Yilin He,Chaojie Wang,Hao Zhang,Bo Chen,Mingyuan Zhou

Graph convolutional networks (GCNs), 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 graph-level and node-level classification tasks. However, GCNs typically treat the graph adjacency matrix as given and ignore how the edges could be formed by different latent inter-node relations. In this paper, we introduce a relational graph generative process to model how the observed edges are generated by aggregating the node interactions over multiple overlapping node communities, each of which represents a particular type of relation that contributes to the edges via a logical OR mechanism. Based on this relational generative model, we partition each edge into the summation of multiple relation-specific weighted edges, and use the weighted edges in each community to define a relation-specific GCN. We introduce a variational inference framework to jointly learn how to partition the edges into different communities and combine relation-specific GCNs for the end classification tasks. Extensive evaluations on real-world datasets have demonstrated the working mechanisms of the edge partition modulated GCNs and their efficacy in learning both node and graph-level representations.

****

LASSO: Latent Sub-spaces Orientation for Domain Generalization
Long Tung Vuong,Trung Quoc Phung,Toan Tran,Anh Tuan Tran,Dinh Phung,Trung Le

To achieve a satisfactory generalization performance on prediction tasks in an unseen domain, existing domain generalization (DG) approaches often rely on the strict assumption of fixed domain-invariant features and common hypotheses learned from a set of training domains. While it is a natural and important premise to ground generalization capacity on the target domain, we argue that this assumption could be overly strict and sub-optimal. It is particularly evident when source domains share little information or the target domains leverages information from selective source domains in a compositional way instead of relying on a unique invariant hypothesis across all source domains. Unlike most existing approaches, instead of constructing a single hypothesis shared among domains, we propose a LAtent Sub-Space Orientation (LASSO) method that explores diverse latent sub-spaces and learning individual hypotheses on those sub-spaces. Moreover, in LASSO, since the latent sub-spaces are formed by the label-informative features captured in source domains, they allow us to project target examples onto appropriate sub-spaces, while preserving crucial label-informative features for the label prediction. Finally, we empirically evaluate our method on several well-known DG benchmarks, where it achieves state-of-the-art results.

****

AdaAug: Learning Class- and Instance-adaptive Data Augmentation Policies
Tsz-Him Cheung,Dit-Yan Yeung

Data augmentation is an effective way to improve the generalization capability of modern deep learning models. However, the underlying augmentation methods mostly rely on handcrafted operations. Moreover, an augmentation policy useful to one dataset may not transfer well to other datasets. Therefore, Automated Data Augmentation (AutoDA) methods, like \textit{AutoAugment} and \textit{Population-based Augmentation}, have been proposed recently to automate the process of searching for optimal augmentation policies. However, the augmentation policies found are not adaptive to the dataset used, hindering the effectiveness of these AutoDA methods. In this paper, we propose a novel AutoDA method called \texttt{AdaAug} to efficiently learn adaptive augmentation policies in a class-dependent and potentially instance-dependent manner. Our experiments show that the adaptive augmentation policies learned by our method transfer well to unseen datasets such as the Oxford Flowers, Oxford-IIT Pets, FGVC Aircraft, and Stanford Cars datasets when compared with other AutoDA baselines. In addition, our method also achieves state-of-the-art performance on the CIFAR-10, CIFAR-100, and SVHN datasets.

****

Unsupervised Semantic Segmentation by Distilling Feature Correspondences

Mark Hamilton,Zhoutong Zhang,Bharath Hariharan,Noah Snavely,William T. Freeman

Unsupervised semantic segmentation aims to discover and localize semantically meaningful categories within image corpora without any form of annotation. To solve this task, algorithms must produce features for every pixel that are both semantically meaningful and compact enough to form distinct clusters. Unlike previous works which achieve this with a single end-to-end framework, we propose to separate feature learning from cluster compactification. Empirically, we show that current unsupervised feature learning frameworks already generate dense features whose correlations are semantically consistent. This observation motivates us to design STEGO ($\textbf{S}$elf-supervised $\textbf{T}$ransformer with $\textbf{E}$nergy-based $\textbf{G}$raph $\textbf{O}$ptimization), a novel framework that distills unsupervised features into high-quality discrete semantic labels. At the core of STEGO is a novel contrastive loss function that encourages features to form compact clusters while preserving their association pattern. STEGO yields a significant improvement over the prior state of the art, on both the CocoStuff ($\textbf{+14 mIoU}$) and Cityscapes ($\textbf{+9 mIoU}$) semantic segmentation challenges.

****

Pruning Compact ConvNets For Efficient Inference

Sayan Ghosh,Karthik Prasad,Xiaoliang Dai,Peizhao Zhang,Bichen Wu,Graham Cormode,Peter Vajda

Neural network pruning is frequently used to compress over-parameterized networks by large amounts, while incurring only marginal drops in generalization performance. However, the impact of pruning on networks that have been highly optimized for efficient inference has not received the same level of attention. In this paper, we analyze the effect of pruning for computer vision, and study state-of-the-art FBNetV3 family of models. We show that model pruning approaches can be used to further optimize networks trained through NAS (Neural Architecture Search). The resulting family of pruned models can consistently obtain better performance than existing FBNetV3 models at the same level of computation, and thus provide state-of-the-art results when trading off between computational complexity and generalization performance on the ImageNet benchmark. In addition to better generalization performance, we also demonstrate that when limited computation resources are available, pruning FBNetV3 models incur only a fraction of GPU-hours involved in running a full-scale NAS (Neural Architecture Search).

****

Label Smoothed Embedding Hypothesis for Out-of-Distribution Detection

Dara Bahri,Heinrich Jiang,Yi Tay,Donald Metzler

Detecting out-of-distribution (OOD) examples is critical in many applications. We propose an unsupervised method to detect OOD samples using a k-NN density estimate with respect to a classification model's intermediate activations on in-distribution samples. We leverage a recent insight about label smoothing, which we call the {\it Label Smoothed Embedding Hypothesis}, and show that one of the implications is that the k-NN density estimator performs better as an OOD detection method both theoretically and empirically when the model is trained with label smoothing. Finally, we show that our proposal outperforms many OOD baselines and we also provide new finite-sample high-probability statistical results for k-NN density estimation's ability to detect OOD examples.

****

Axiomatic Explanations for Visual Search, Retrieval, and Similarity Learning

Mark Hamilton,Scott Lundberg,Stephanie Fu,Lei Zhang,William T. Freeman

Visual search, recommendation, and contrastive similarity learning power technologies that impact billions of users worldwide. Modern model architectures can be complex and difficult to interpret, and there are several competing techniques one can use to explain a search engine's behavior. We show that the theory of fair credit assignment provides a unique axiomatic solution that generalizes several existing recommendation- and metric-explainability techniques in the literature. Using this formalism, we show when existing approaches violate "fairness" and derive methods that sidestep these shortcomings and naturally handle counterfactual information. More specifically, we show existing approaches implicitly app roximate second-order Shapley-Taylor indices and extend CAM, GradCAM, LIME, SHAP, SBSM, and other methods to search engines. These extensions can extract pairwise correspondences between images from trained opaque-box models. We also introduce a fast kernel-based method for estimating Shapley-Taylor indices that require orders of magnitude fewer function evaluations to converge. Finally, we show that these game-theoretic measures yield more consistent explanations for image similarity architectures.

****

On the Safety of Interpretable Machine Learning: A Maximum Deviation Approach

Dennis Wei,Rahul Nair,Amit Dhurandhar,Kush R. Varshney,Elizabeth M. Daly,Moninder Singh

Interpretable and explainable machine learning has seen a recent surge of interest. We posit that safety is a key reason behind the demand for explainability. To explore this relationship, we propose a mathematical formulation for assessing the safety of supervised learning models based on their maximum deviation over a certification set. We then show that for interpretable models including decision trees, rule lists, 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 repurpose results from the multi-armed bandit literature to show that interpretability produces tighter (regret) bounds on the maximum deviation compared with black box functions. We perform experiments that quantify the dependence of the maximum deviation on model smoothness and certification set size. The experiments also illustrate how the solutions that maximize deviation can suggest safety risks.

****

Graph-Relational Domain Adaptation

Zihao Xu,Hao He,Guang-He Lee,Bernie Wang,Hao Wang

Existing domain adaptation methods tend to treat every domain equally and align them all perfectly. Such uniform alignment ignores topological structures among different domains; therefore it may be beneficial for nearby domains, but not necessarily for distant domains. In this work, we relax such uniform alignment by using a domain graph to encode domain adjacency, e.g., a graph of states in the US with each state as a domain and each edge indicating adjacency, thereby allowing domains to align flexibly based on the graph structure. We generalize the existing adversarial learning framework with a novel graph discriminator using encoding-conditioned graph embeddings. Theoretical analysis shows that at equilibrium, our method recovers classic domain adaptation when the graph is a clique, and achieves non-trivial alignment for other types of graphs. Empirical results show that our approach successfully generalizes uniform alignment, naturally incorporates domain information represented by graphs, and improves upon existing domain adaptation methods on both synthetic and real-world datasets.

****

Training invariances and the low-rank phenomenon: beyond linear networks

Thien Le,Stefanie Jegelka

The implicit bias induced by the training of neural networks has become a topic of rigorous study. In the limit of gradient flow and gradient descent with appropriate step size, it has been shown that when one trains a deep linear network with logistic or exponential loss on linearly separable data, the weights converge to rank-1 matrices. In this paper, we extend this theoretical result to the last few linear layers of the much wider class of nonlinear ReLU-activated feedforward networks containing fully-connected layers and skip connections. Similar to the linear case, the proof relies on specific local training invariances, sometimes referred to as alignment, which we show to hold for submatrices where neurons are stably-activated in all training examples, and it reflects empirical results in the literature. We also show this is not true in general for the full matrix of ReLU fully-connected layers. Our proof relies on a specific decomposition of the network into a multilinear function and another ReLU network whose weights are constant under a certain parameter directional convergence.

****

Generating Antimicrobial Peptides from Latent Secondary Structure Space

Danqing Wang,Zeyu Wen,Lei Li,Hao Zhou

Antimicrobial peptides (AMPs) have shown promising results in broad-spectrum antibiotics and resistant infection treatments, which makes it attract plenty of attention in drug discovery. Recently, many researchers bring deep generative models to AMP design. However, few studies consider structure information during the generation, though it has shown crucial influence on antimicrobial activity in all AMP mechanism theories. In this paper, we propose LSSAMP that uses the multi-scale VQ-VAE to learn the positional latent spaces modeling the secondary structure. By sampling in the latent secondary structure space, we can generate peptides with ideal amino acids and secondary structures at the same time. Experimental results show that our LSSAMP can generate peptides with multiply ideal physical attributes and a high probability of being predicted as AMPs by public AMP prediction models.

****

Learning Long-Term Reward Redistribution via Randomized Return Decomposition

Zhizhou Ren,Ruihan Guo,Yuan Zhou,Jian Peng

Many practical applications of reinforcement learning require agents to learn from sparse and delayed rewards. It challenges the ability of agents to attribute their actions to future outcomes. In this paper, we consider the problem formulation of episodic reinforcement learning with trajectory feedback. It refers to an extreme delay of reward signals, in which the agent can only obtain one reward signal at the end of each trajectory. A popular paradigm for this problem setting is learning with a designed auxiliary dense reward function, namely proxy reward, instead of sparse environmental signals. Based on this framework, this paper proposes a novel reward redistribution algorithm, randomized return decomposition (RRD), to learn a proxy reward function for episodic reinforcement learning. We establish a surrogate problem by Monte-Carlo sampling that scales up least-squares-based reward redistribution to long-horizon problems. We analyze our surrogate loss function by connection with existing methods in the literature, which illustrates the algorithmic properties of our approach. In experiments, we extensively evaluate our proposed method on a variety of benchmark tasks with episodic rewards and demonstrate substantial improvement over baseline algorithms.

****

Improving Meta-Continual Learning Representations with Representation Replay

Lawrence Ki-On Chan,James Kwok

Continual learning often suffers from catastrophic forgetting. Recently, meta-continual learning algorithms use meta-learning to learn how to continually learn. A recent state-of-the-art is online aware meta-learning (OML). This can be further improved by incorporating experience replay (ER) into its meta-testing. However, the use of ER only in meta-testing but not in meta-training suggests that the model may not be optimally meta-trained. In this paper, we remove this inconsistency in the use of ER and improve continual learning representations by integrating ER also into meta-training. We propose to store the samples' representations, instead of the samples themselves, into the replay buffer. This ensures the batch nature of ER does not conflict with the online-aware nature of OML. Moreover, we introduce a meta-learned sample selection scheme to replace the widely used reservoir sampling to populate the replay buffer. This allows the most significant samples to be stored, rather than relying on randomness. Class-balanced modifiers are further added to the sample selection scheme to ensure each class has sufficient samples stored in the replay buffer. Experimental results on a number of real-world meta-continual learning benchmark data sets demonstrate that the proposed method outperforms the state-of-the-art. Moreover, the learned representations have better clustering structures and are more discriminative.

****

Indiscriminate Poisoning Attacks Are Shortcuts

Da Yu,Huishuai Zhang,Wei Chen,Jian Yin,Tie-Yan Liu

Indiscriminate data poisoning attacks, which add imperceptible perturbations to training data to maximize the test error of trained models, have become a trendy topic because they are thought to be capable of preventing unauthorized use of

data. In this work, we investigate why these perturbations work in principle. We find that the perturbations of advanced poisoning attacks are almost linear separable when assigned with the target labels of the corresponding samples. This is an important population property for various perturbations that were not unveiled before. Moreover, we further confirm that linear separability is indeed the workhorse for poisoning attacks. We synthesize linear separable data as perturbations and show that such synthetic perturbations are as powerful as the deliberately crafted attacks. Our finding also suggests that the shortcut learning problem is more serious than previously believed as deep learning heavily relies on shortcuts even if they are of an imperceptible scale and mixed together with the normal features. It also suggests that pre-trained feature extractors can be a powerful defense.
****
Revisit Kernel Pruning with Lottery Regulated Grouped Convolutions
Shaochen Zhong,Guanqun Zhang,Ningjia Huang,Shuai Xu
Structured pruning methods which are capable of delivering a densely pruned network are among the most popular techniques in the realm of neural network pruning, where most methods prune the original network at a filter or layer level. Although such methods may provide immediate compression and acceleration benefits, we argue that the blanket removal of an entire filter or layer may result in undesired accuracy loss. In this paper, we revisit the idea of kernel pruning (to only prune one or several $k \times k$ kernels out of a 3D-filter), a heavily overlooked approach under the context of structured pruning. This is because kernel pruning will naturally introduce sparsity to filters within the same convolutional layer — thus, making the remaining network no longer dense. We address this problem by proposing a versatile grouped pruning framework where we first cluster filters from each convolutional layer into equal-sized groups, prune the grouped kernels we deem unimportant from each filter group, then permute the remaining filters to form a densely grouped convolutional architecture (which also enables the parallel computing capability) for fine-tuning. Specifically, we consult empirical findings from a series of literature regarding $\textit{Lottery Ticket Hypothesis}$ to determine the optimal clustering scheme per layer, and develop a simple yet cost-efficient greedy approximation algorithm to determine which group kernels to keep within each filter group. Extensive experiments also demonstrate our method often outperforms comparable SOTA methods with lesser data augmentation needed, smaller fine-tuning budget required, and sometimes even much simpler procedure executed (e.g., one-shot v. iterative). Please refer to our GitHub repository (https://github.com/choH/lottery_regulated_grouped_kernel_pruning) for code.
****
Bi-linear Value Networks for Multi-goal Reinforcement Learning
Zhang-Wei Hong,Ge Yang,Pulkit Agrawal
Universal value functions are a core component of off-policy multi-goal reinforcement learning.
The de-facto paradigm is to approximate Q(s, a, g) using monolithic neural networks which lack inductive biases to produce complex interactions between the state s and the goal g. In this work, we propose a bilinear decomposition that represents the Q-value via a low-rank approximation in the form of a dot product between two vector fields. The first vector field, f(s, a), captures the environment's local dynamics at the state s; whereas the second component, φ(s, g), captures the global relationship between the current state and the goal.
We show that our bilinear decomposition scheme improves sample efficiency over the original monolithic value approximators, and transfer better to unseen goals. We demonstrate significant learning speed-up over a variety of tasks on a simulated robot arm, and the challenging task of dexterous manipulation with a Shadow hand.
****
PIVQGAN: Posture and Identity Disentangled Image-to-Image Translation via Vector Quantization
Bingchen Liu,Yizhe Zhu,Xiao Yang,Ahmed Elgammal

One popular objective for the image-to-image translation task is to independently control the coarse-level object arrangements (posture) and the fine-grained level styling (identity) of the generated image from two exemplar sources. To approach this objective, we propose PIVQGAN with two novel techniques in the framework of StyleGAN2. First, we propose a Vector-Quantized Spatial Normalization (VQSN) module for the generator for better pose-identity disentanglement. The VQSN module automatically learns to encode the shaping and composition information from the commonly shared objects inside the training-set images. Second, we design a joint-training scheme with self-supervision methods for the GAN-Inversion encoder and the generator. Specifically, we let the encoder and generator reconstruct images from two differently augmented variants of the original ones, one defining the pose and the other for identity. The VQSN module facilitates a more delicate separation of posture and identity, while the training scheme ensures the VQSN module learns the pose-related representations. Comprehensive experiments conducted on various datasets show better synthesis image quality and disentangling scores of our model. Moreover, we present model applications beyond posture-identity disentangling, thanks to the latent-space reducing feature of the leveraged VQSN module.

****

SpSC: A Fast and Provable Algorithm for Sampling-Based GNN Training
Shihui Song,Peng Jiang

Neighbor sampling is a commonly used technique for training Graph Neural Networks (GNNs) on large graphs. Previous work has shown that sampling-based GNN training can be considered as Stochastic Compositional Optimization (SCO) problems and can be better solved by SCO algorithms. However, we find that SCO algorithms are impractical for training GNNs on large graphs because they need to store the moving averages of the aggregated features of all nodes in the graph. The moving averages can easily exceed the GPU memory limit and even the CPU memory limit. In this work, we propose a variant of SCO algorithms with sparse moving averages for GNN training. By storing the moving averages in the most recent iterations, our algorithm only requires a fixed size buffer, regardless of the graph size. We show that our algorithm can achieve $O(\sqrt{1/K})$ convergence rate when the buffer size satisfies certain conditions. Our experiments validate our theoretical results and show that our algorithm outperforms the traditional Adam SGD for GNN training with a small memory overhead.

****

No One Representation to Rule Them All: Overlapping Features of Training Methods
Raphael Gontijo-Lopes,Yann Dauphin,Ekin Dogus Cubuk

Despite being able to capture a range of features of the data, high accuracy models trained with supervision tend to make similar predictions. This seemingly implies that high-performing models share similar biases regardless of training methodology, which would limit ensembling benefits and render low-accuracy models as having little practical use. Against this backdrop, recent work has developed quite different training techniques, such as large-scale contrastive learning, yielding competitively high accuracy on generalization and robustness benchmarks. This motivates us to revisit the assumption that models necessarily learn similar functions. We conduct a large-scale empirical study of models across hyper-parameters, architectures, frameworks, and datasets. We find that model pairs that diverge more in training methodology display categorically different generalization behavior, producing increasingly uncorrelated errors. We show these models specialize in subdomains of the data, leading to higher ensemble performance: with just 2 models (each with ImageNet accuracy \~76.5\%), we can create ensembles with 83.4\% (+7\% boost). Surprisingly, we find that even significantly low-accuracy models can be used to improve high-accuracy models. Finally, we show diverging training methodology yield representations that capture overlapping (but not supersetting) feature sets which, when combined, lead to increased downstream performance.

****

Neural Structure Mapping For Learning Abstract Visual Analogies
Shashank Shekhar,Graham W. Taylor

Building conceptual abstractions from sensory information and then reasoning about them is central to human intelligence. Abstract reasoning both relies on, and is facilitated by, our ability to make analogies about concepts from known domains to novel domains. Structure Mapping Theory of human analogical reasoning posits that analogical mappings rely on (higher-order) relations and not on the sensory content of the domain. This enables humans to reason systematically about novel domains, a problem with which machine learning (ML) models tend to struggle. We introduce a two-stage neural framework, which we label Neural Structure Mapping (NSM), to learn visual analogies from Raven's Progressive Matrices, an abstract visual reasoning test of fluid intelligence. Our framework uses (1) a multi-task visual relationship encoder to extract constituent concepts from raw visual input in the source domain, and (2) a neural module net analogy inference engine to reason compositionally about the inferred relation in the target domain. Our NSM approach (a) isolates the relational structure from the source domain with high accuracy, and (b) successfully utilizes this structure for analogical reasoning in the target domain.
****

What Happens after SGD Reaches Zero Loss? --A Mathematical Framework
Zhiyuan Li,Tianhao Wang,Sanjeev Arora

Understanding the implicit bias of Stochastic Gradient Descent (SGD) is one of the key challenges in deep learning, especially for overparametrized models, where the local minimizers of the loss function L can form a manifold. Intuitively, with a sufficiently small learning rate η, SGD tracks Gradient Descent (GD) until it gets close to such manifold, where the gradient noise prevents further convergence. In such regime, Blanc et al. (2020) proved that SGD with label noise locally decreases a regularizer-like term, the sharpness of loss, $\text{tr}[\nabla^2 L]$. The current paper gives a general framework for such analysis by adapting ideas from Katzenberger (1991). It allows in principle a complete characterization for the regularization effect of SGD around such manifold---i.e., the "implicit bias"---using a stochastic differential equation (SDE) describing the limiting dynamics of the parameters, which is determined jointly by the loss function and the noise covariance. This yields some new results: (1) a *global* analysis of the implicit bias valid for η^{-2} steps, in contrast to the local analysis of Blanc et al. (2020) that is only valid for $\eta^{-1.6}$ steps and (2) allowing *arbitrary* noise covariance. As an application, we show with arbitrary large initialization, label noise SGD can always escape the kernel regime and only requires $O(\kappa\ln d)$ samples for learning an κ-sparse overparametrized linear model in $\mathbb{R}^d$ (Woodworth et al., 2020), while GD initialized in the kernel regime requires $\Omega(d)$ samples. This upper bound is minimax optimal and improves the previous $\widetilde{O}(\kappa^2)$ upper bound (HaoChen et al., 2020).
****

How does Contrastive Pre-training Connect Disparate Domains?
Kendrick Shen,Robbie Matthew Jones,Ananya Kumar,Sang Michael Xie,Percy Liang

Pre-training on massive unlabeled datasets greatly improves accuracy under distribution shifts. As a first step toward understanding this, we study a popular pre-training method, contrastive learning, in the unsupervised domain adaptation (UDA) setting where we only have labeled data from a source domain and unlabeled data from a target domain. We begin by showing on 4 benchmark datasets that out-of-the-box contrastive pre-training (even without large-scale unlabeled data) is competitive with other UDA methods. Intuitions from classical UDA methods such as domain adversarial training focus on bringing the domains together in feature space to improve generalization from source to target. Surprisingly, we find that contrastive pre-training learns features that are very far apart between the source and target domains. How then does contrastive learning improve robustness to distribution shift? We develop a conceptual model for contrastive learning under domain shifts, where data augmentations form connections between classes and domains that can be far apart. We propose a new measure of connectivity ---the relative connection strengths between same and different classes across domains ---that governs the success of contrastive pre-training for domain adaptation in a simple example and strongly correlates with our results on benchmark datasets.

****

GCF: Generalized Causal Forest for Heterogeneous Treatment Effect Estimation Using Nonparametric Methods

Shu Wan,Chen Zheng,Zhonggen Sun,Mengfan Xu,Xiaoqing Yang,Jiecheng Guo,Hongtu Zhu

Heterogeneous treatment effect (HTE) estimation with continuous treatment is essential in multiple disciplines, such as the online marketplace and pharmaceutical industry. The existing machine learning (ML) methods, like forest-based modeling, either work only for discrete treatments or make partially linear or parametric assumptions that may suffer from model misspecification. To alleviate these problems, we extend causal forest (CF) with non-parametric dose-response functions (DRFs) that can be estimated locally using kernel-based Double/Debiased ML estimators. Moreover, we propose a distance-based splitting criterion in the functional space of Partial DRFs to capture the heterogeneity for continuous treatments. We call the proposed algorithm generalized causal forest (GCF) as it generalizes the use case of CF to a much broader setup. We show the effectiveness of GCF compared to SOTA on synthetic data and proprietary real-world data sets.

****

One Objective for All Models --- Self-supervised Learning for Topic Models

Zeping Luo,Cindy Weng,Shiyou Wu,Mo Zhou,Rong Ge

Self-supervised learning has significantly improved the performance of many NLP tasks. In this paper, we highlight a key advantage of self-supervised learning - when applied to data generated by topic models, self-supervised learning can be oblivious to the specific model, and hence is less susceptible to model mis-specification. In particular, we prove that commonly used self-supervised objectives based on reconstruction or contrastive samples can both recover useful posterior information for general topic models. Empirically, we show that the same objectives can perform competitively against posterior inference using the correct model, while outperforming posterior inference using mis-specified model.

****

ShiftAddNAS: Hardware-Inspired Search for More Accurate and Efficient Neural Networks

Haoran You,Baopu Li,Huihong Shi,Yingyan Lin

Neural networks (NNs) with intensive multiplications (e.g., convolutions and transformers) are powerful yet power hungry, impeding their more extensive deployment into resource-constrained edge devices. As such, multiplication-free networks, which follow a common practice in energy-efficient hardware implementation to parameterize NNs with more efficient operators (e.g., bitwise shifts and additions), have gained growing attention. However, multiplication-free networks in general under-perform their vanilla counterparts in terms of the achieved accuracy. To this end, this work advocates hybrid NNs that consist of both powerful yet costly multiplications and efficient yet less powerful operators for marrying the best of both worlds, and proposes ShiftAddNAS, which can automatically search for more accurate and more efficient NNs. Our ShiftAddNAS highlights two enablers. Specifically, it integrates (1) the first hybrid search space that incorporates both multiplication-based and multiplication-free operators for facilitating the development of both accurate and efficient hybrid NNs; and (2) a novel weight sharing strategy that enables effective weight sharing among different operators that follow heterogeneous distributions (e.g., Gaussian for convolutions vs. Laplacian for add operators) and simultaneously leads to a largely reduced supernet size and much better searched networks. Extensive experiments and ablation studies on various models, datasets, and tasks consistently validate the effectiveness of ShiftAddNAS, e.g., achieving up to a +7.7% higher accuracy or a +4.9 better BLEU score as compared to state-of-the-art expert-designed and neural architecture searched NNs, while leading to up to 93% or 69% energy and latency savings, respectively. All the codes will be released upon acceptance.

****

Generalized Kernel Thinning

Raaz Dwivedi,Lester Mackey

The kernel thinning (KT) algorithm of Dwivedi and Mackey (2021) compresses a probability distribution more effectively than independent sampling by targeting a reproducing kernel Hilbert space (RKHS) and leveraging a less smooth square-root kernel. Here we provide four improvements. First, we show that KT applied directly to the target RKHS yields tighter, dimension-free guarantees for any kernel, any distribution, and any fixed function in the RKHS. Second, we show that, for analytic kernels like Gaussian, inverse multiquadric, and sinc, target KT admits maximum mean discrepancy (MMD) guarantees comparable to or better than those of square-root KT without making explicit use of a square-root kernel. Third, we prove that KT with a fractional power kernel yields better-than-Monte-Carlo MMD guarantees for non-smooth kernels, like Laplace and Matern, that do not have square-roots. Fourth, we establish that KT applied to a sum of the target and power kernels (a procedure we call KT+) simultaneously inherits the improved MMD guarantees of power KT and the tighter individual function guarantees of target KT. In our experiments with target KT and KT+, we witness significant improvements in integration error even in 100 dimensions and when compressing challenging differential equation posteriors.

****

How Much Can CLIP Benefit Vision-and-Language Tasks?
Sheng Shen,Liunian Harold Li,Hao Tan,Mohit Bansal,Anna Rohrbach,Kai-Wei Chang,Zhewei Yao,Kurt Keutzer

Most existing Vision-and-Language (V&L) models rely on pre-trained visual encoders, using a relatively small set of manually-annotated data (as compared to web-crawled data), to perceive the visual world. However, it has been observed that large-scale pretraining usually can result in better generalization performance, e.g., CLIP (Contrastive Language-Image Pre-training), trained on a massive amount of image-caption pairs, has shown a strong zero-shot capability on various vision tasks. To further study the advantage brought by CLIP, we propose to use CLIP as the visual encoder in various V&L models in two typical scenarios: 1) plugging CLIP into task-specific fine-tuning; 2) combining CLIP with V&L pre-training and transferring to downstream tasks. We show that CLIP significantly outperforms widely-used visual encoders trained with in-domain annotated data, such as BottomUp-TopDown. We achieve competitive or better results on diverse V&L tasks, while establishing new state-of-the-art results on Visual Question Answering, Visual Entailment, and V&L Navigation tasks.

****

Large Learning Rate Tames Homogeneity: Convergence and Balancing Effect
Yuqing Wang,Minshuo Chen,Tuo Zhao,Molei Tao

Recent empirical advances show that training deep models with large learning rate often improves generalization performance. However, theoretical justifications on the benefits of large learning rate are highly limited, due to challenges in analysis. In this paper, we consider using Gradient Descent (GD) with a large learning rate on a homogeneous matrix factorization problem, i.e., $\min_{X, Y} \|A - XY^\top\|_{\sf F}^2$. We prove a convergence theory for constant large learning rates well beyond $2/L$, where L is the largest eigenvalue of Hessian at the initialization. Moreover, we rigorously establish an implicit bias of GD induced by such a large learning rate, termed `balancing', meaning that magnitudes of X and Y at the limit of GD iterations will be close even if their initialization is significantly unbalanced. Numerical experiments are provided to support our theory.

****

Demystifying Limited Adversarial Transferability in Automatic Speech Recognition Systems
Hadi Abdullah,Aditya Karlekar,Vincent Bindschaedler,Patrick Traynor

The targeted transferability of adversarial samples enables attackers to exploit black-box models in the real-world. The most popular method to produce these adversarial samples is optimization attacks, which have been shown to achieve a high level of transferability in some domains. However, recent research has demonstrated that these attack samples fail to transfer when applied to Automatic Spee ch Recognition Systems (ASRs). In this paper, we investigate factors preventing this transferability via exhaustive experimentation. To do so, we perform an ablation study on each stage of the ASR pipeline. We discover and quantify six factors (i.e., input type, MFCC, RNN, output type, and vocabulary and sequence sizes) that impact the targeted transferability of optimization attacks against ASRs. Future research can leverage our findings to build ASRs that are more robust to other transferable attack types (e.g., signal processing attacks), or to modify architectures in other domains to reduce their exposure to targeted transferability of optimization attacks.

****

Learning to perceive objects by prediction

Tushar Arora,Li Erran Li,Ming Bo Cai

The representation of objects is the building block of higher-level concepts. Infants develop the notion of objects without supervision. The prediction error of future sensory input is likely the major teaching signal for infants. Inspired by this, we propose a new framework to extract object-centric representation from single 2D images by learning to predict future scenes in the presence of moving objects. We treat objects as latent causes whose function to an agent is to facilitate efficient prediction of the coherent motion of their parts in visual input. Distinct from previous object-centric models, our model learn to explicitly infer objects' location in 3D environment in addition to segmenting objects. Further, the network learns a latent code space where objects with the same geometric shape and texture/color frequently group together. The model requires no supervision or pre-training of any part of the network. We provide a new synthetic dataset with more complex textures on objects and background and found several previous models not based on predictive learning overly rely on clustering colors and lose specificity in object segmentation. Our work demonstrates a new approach for learning symbolic representation grounded in sensation and action.

****

IA-MARL: Imputation Assisted Multi-Agent Reinforcement Learning for Missing Training Data

Dongsun Kim,Sinwoong Yun,Jemin Lee,Eunbyung Park

Recently, multi-agent reinforcement learning (MARL) adopts the centralized training with decentralized execution (CTDE) framework that trains agents using the data from all agents at a centralized server while each agent takes an action from its observation. In the real world, however, the training data from some agents can be unavailable at the centralized server due to practical reasons including communication failures and security attacks (e.g., data modification), which can slow down training and harm performance. Therefore, we consider the missing training data problem in MARL, and then propose the imputation assisted multiagent reinforcement learning (IA-MARL). IA-MARL consists of two steps: 1) the imputation of missing training data, which uses generative adversarial imputation networks (GAIN), and 2) the mask-based update of the networks, which trains each agent using the training data of corresponding agent, not missed over consecutive times. In the experimental results, we explore the effects of the data missing probability, the number of agents, and the number of pre-training episodes for GAIN on the performance of IA-MARL. We show IA-MARL outperforms a decentralized approach and even can achieve the performance of MARL without missing training data when sufficient imputation accuracy is supported. Our ablation study also shows that both the mask-based update and the imputation accuracy play important roles in achieving the high performance in IA-MARL.

****

ES-Based Jacobian Enables Faster Bilevel Optimization

Daouda Sow,Kaiyi Ji,Yingbin Liang

Bilevel optimization (BO) has arisen as a powerful tool for solving many modern machine learning problems. However, due to the nested structure of BO, existing gradient-based methods require second-order derivative approximations via Jacobian- or/and Hessian-vector computations, which can be very costly in practice, especially with large neural network models. In this work, we propose a novel BO algorithm, which adopts Evolution Strategies (ES) based method to approximate the response Jacobian matrix in the hypergradient of BO, and hence fully eliminates all second-order computations. We call our algorithm as ESJ (which stands for the ES-based Jacobian method) and further extend it to the stochastic setting as ESJ-S. Theoretically, we characterize the convergence guarantee and computational complexity for our algorithms. Experimentally, we demonstrate the superiority of our proposed algorithms compared to the state of the art methods on various bilevel problems. Particularly, in our experiment in the few-shot meta-learning problem, we meta-learn the twelve millions parameters of a ResNet-12 network over the miniImageNet dataset, which evidently demonstrates the scalability of our ES-based bilevel approach and its feasibility in the large-scale setting.

****

PipeGCN: Efficient Full-Graph Training of Graph Convolutional Networks with Pipelined Feature Communication

Cheng Wan,Youjie Li,Cameron R. Wolfe,Anastasios Kyrillidis,Nam Sung Kim,Yingyan Lin

Graph Convolutional Networks (GCNs) is the state-of-the-art method for learning graph-structured data, and training large-scale GCNs requires distributed training across multiple accelerators such that each accelerator is able to hold a partitioned subgraph. However, distributed GCN training incurs prohibitive overhead of communicating node features and feature gradients among partitions for every GCN layer during each training iteration, limiting the achievable training efficiency and model scalability. To this end, we propose PipeGCN, a simple yet effective scheme that hides the communication overhead by pipelining inter-partition communication with intra-partition computation. It is non-trivial to pipeline for efficient GCN training, as communicated node features/gradients will become stale and thus can harm the convergence, negating the pipeline benefit. Notably, little is known regarding the convergence rate of GCN training with both stale features and stale feature gradients. This work not only provides a theoretical convergence analysis but also finds the convergence rate of PipeGCN to be close to that of the vanilla distributed GCN training without any staleness. Furthermore, we develop a smoothing method to further improve PipeGCN's convergence. Extensive experiments show that PipeGCN can largely boost the training throughput (1.7×~28.5×) while achieving the same accuracy as its vanilla counterpart and existing full-graph training methods. The code is available at https://github.com/RICE-EIC/PipeGCN.

****

Learning Neural Contextual Bandits through Perturbed Rewards

Yiling Jia,Weitong ZHANG,Dongruo Zhou,Quanquan Gu,Hongning Wang

Thanks to the power of representation learning, neural contextual bandit algorithms demonstrate remarkable performance improvement against their classical counterparts. But because their exploration has to be performed in the entire neural network parameter space to obtain nearly optimal regret, the resulting computational cost is prohibitively high.

We propose to perturb the rewards when updating the neural network to eliminate the need of explicit exploration and the corresponding computational overhead. We prove that a $\tilde{O}(\tilde{d}\sqrt{T})$ regret upper bound is still achievable under standard regularity conditions, where T is the number of rounds of interactions and $\tilde{d}$ is the effective dimension of a neural tangent kernel matrix.

Extensive comparisons with several benchmark contextual bandit algorithms, including two recent neural contextual bandit models, demonstrate the effectiveness and computational efficiency of our proposed neural bandit algorithm.

****

Knowledge Guided Geometric Editing for Unsupervised Drug Design

Yuwei Yang,Siqi Ouyang,Meihua Dang,Mingyue Zheng,Lei Li,Hao Zhou

Deep learning models have been widely used in automatic drug design. Current deep approaches always represent and generate candidate molecules as a 1D string or a 2D graph, which rely on large measurement data from lab experiments for training. However, many disease targets in particular newly discovered ones do not have such data available. In this paper, we propose \method, which incorporates p hysicochemical knowledge into deep models, leading to unsupervised drug design. Specifically, \method directly models drug molecules in the geometric~(3D) space and performs geometric editing with the knowledge guidance by self-training and simulated annealing in a purely training data free fashion. Our experimental results demonstrate that GEKO outperforms baselines on all 12 targets with and without prior drug-target measurement data.

****

Graph Information Matters: Understanding Graph Filters from Interaction Probability

Zhixian Chen,Tengfei Ma,Yang Wang

Graph Neural Networks (GNNs) have received extensive affirmation for their promising performance in graph learning problems. Despite their various neural architectures, most are intrinsically graph filters that provide theoretical foundations for model explanations. In particular, low-pass filters show superiority in label prediction in many benchmarks. However, recent empirical research suggests that models with only low pass filters do not always perform well. Although increasing attempts to understand graph filters, it is unclear how a particular graph affects the performance of different filters. In this paper, we carry out a comprehensive theoretical analysis of the synergy of graph structure and node features on graph filters’ behaviors in node classification, relying on the introduction of interaction probability and frequency distribution. We show that the homophily degree of graphs significantly affects the prediction error of graph filters. Our theory provides a guideline for graph filter design in a data-driven manner. Since it is hard for a single graph filter to live up to this, we propose a general strategy for exploring a data-specified filter bank. Experimental results show that our model achieves consistent and significant performance improvements across all benchmarks. Furthermore, we empirically validate our theoretical analysis and explain the behavior of baselines and our model.

****

Feature Selection in the Contrastive Analysis Setting

Ethan Weinberger,Ian Connick Covert,Su-In Lee

The goal of unsupervised feature selection is to select a small number of informative features for use in unknown downstream tasks. Here the definition of ``informative'' is subjective and dependent on the specifics of a given problem domain. In the contrastive analysis (CA) setting, machine learning practitioners are specifically interested in discovering patterns that are enriched in a target dataset as compared to a background dataset generated from sources of variation irrelevant to the task at hand. For example, a biomedical data analyst may wish to find a small set of genes to use as a proxy for variations in genomic data only present among patients with a given disease as opposed to healthy control subjects. However, as of yet the problem of unsupervised feature selection in the CA setting has received little attention from the machine learning community. In this work we present CFS (Contrastive Feature Selection), a method for performing feature selection in the CA setting. We experiment with multiple variations of our method on a semi-synthetic dataset and four real-world biomedical datasets, and we find that it consistently outperforms previous state-of-the-art methods designed for standard unsupervised feature selection scenarios.

****

Graph-Augmented Normalizing Flows for Anomaly Detection of Multiple Time Series

Enyan Dai,Jie Chen

Anomaly detection is a widely studied task for a broad variety of data types; among them, multiple time series appear frequently in applications, including for example, power grids and traffic networks. Detecting anomalies for multiple time series, however, is a challenging subject, owing to the intricate interdependencies among the constituent series. We hypothesize that anomalies occur in low density regions of a distribution and explore the use of normalizing flows for unsupervised anomaly detection, because of their superior quality in density estimation. Moreover, we propose a novel flow model by imposing a Bayesian network among constituent series. A Bayesian network is a directed acyclic graph (DAG) that models causal relationships; it factorizes the joint probability of the series

into the product of easy-to-evaluate conditional probabilities. We call such a graph-augmented normalizing flow approach GANF and propose joint estimation of the DAG with flow parameters. We conduct extensive experiments on real-world datasets and demonstrate the effectiveness of GANF for density estimation, anomaly detection, and identification of time series distribution drift.
****

Efficiently Modeling Long Sequences with Structured State Spaces
Albert Gu,Karan Goel,Christopher Re

A central goal of sequence modeling is designing a single principled model that can address sequence data across a range of modalities and tasks, particularly on long-range dependencies. Although conventional models including RNNs, CNNs, and Transformers have specialized variants for capturing long dependencies, they still struggle to scale to very long sequences of 10000 or more steps. A promising recent approach proposed modeling sequences by simulating the fundamental state space model (SSM) \(x'(t) = Ax(t) + Bu(t), y(t) = Cx(t) + Du(t) \), and showed that for appropriate choices of the state matrix \(A \), this system could handle long-range dependencies mathematically and empirically. However, this method has prohibitive computation and memory requirements, rendering it infeasible as a general sequence modeling solution. We propose the Structured State Space sequence model (S4) based on a new parameterization for the SSM, and show that it can be computed much more efficiently than prior approaches while preserving their theoretical strengths. Our technique involves conditioning \(A \) with a low-rank correction, allowing it to be diagonalized stably and reducing the SSM to the well-studied computation of a Cauchy kernel. S4 achieves strong empirical results across a diverse range of established benchmarks, including (i) 91\% accuracy on sequential CIFAR-10 with no data augmentation or auxiliary losses, on par with a larger 2-D ResNet, (ii) substantially closing the gap to Transformers on image and language modeling tasks, while performing generation $60\times$ faster (iii) SoTA on every task from the Long Range Arena benchmark, including solving the challenging Path-X task of length 16k that all prior work fails on, while being as efficient as all competitors.
****

Adversarial Unlearning of Backdoors via Implicit Hypergradient
Yi Zeng,Si Chen,Won Park,Zhuoqing Mao,Ming Jin,Ruoxi Jia

We propose a minimax formulation for removing backdoors from a given poisoned model based on a small set of clean data. This formulation encompasses much of prior work on backdoor removal. We propose the Implicit Backdoor Adversarial Unlearning (I-BAU) algorithm to solve the minimax. Unlike previous work, which breaks down the minimax into separate inner and outer problems, our algorithm utilizes the implicit hypergradient to account for the interdependence between inner and outer optimization. We theoretically analyze its convergence and the generalizability of the robustness gained by solving minimax on clean data to unseen test data. In our evaluation, we compare I-BAU with six state-of-art backdoor defenses on eleven backdoor attacks over two datasets and various attack settings, including the common setting where the attacker targets one class as well as important but underexplored settings where multiple classes are targeted. I-BAU's performance is comparable to and most often significantly better than the best baseline. Particularly, its performance is more robust to the variation on triggers, attack settings, poison ratio, and clean data size. Moreover, I-BAU requires less computation to take effect; particularly, it is more than $13\times$ faster than the most efficient baseline in the single-target attack setting. Furthermore, it can remain effective in the extreme case where the defender can only access 100 clean samples---a setting where all the baselines fail to produce acceptable results.
****

Maximizing Ensemble Diversity in Deep Reinforcement Learning
Hassam Sheikh,Mariano Phielipp,Ladislau Boloni

Modern deep reinforcement learning (DRL) has been successful in solving a range of challenging sequential decision-making problems. Most of these algorithms use an ensemble of neural networks as their backbone structure and benefit from the diversity among the neural networks to achieve optimal results. Unfortunately, the members of the ensemble can converge to the same point either the parametric space or representation space during the training phase, therefore, losing all the leverage of an ensemble. In this paper, we describe Maximize Ensemble Diversity in Reinforcement Learning (MED-RL), a set of regularization methods inspired from the economics and consensus optimization to improve diversity in the ensemble-based deep reinforcement learning methods by encouraging inequality between the networks during training. We integrated MED-RL in five of the most common ensemble-based deep RL algorithms for both continuous and discrete control tasks and evaluated on six Mujoco environments and six Atari games. Our results show that MED-RL augmented algorithms outperform their un-regularized counterparts significantly and in some cases achieved more than 300$\%$ in performance gains.

****

Graph Neural Networks with Learnable Structural and Positional Representations
Vijay Prakash Dwivedi,Anh Tuan Luu,Thomas Laurent,Yoshua Bengio,Xavier Bresson

Graph neural networks (GNNs) have become the standard learning architectures for graphs. GNNs have been applied to numerous domains ranging from quantum chemistry, recommender systems to knowledge graphs and natural language processing. A major issue with arbitrary graphs is the absence of canonical positional information of nodes, which decreases the representation power of GNNs to distinguish e.g. isomorphic nodes and other graph symmetries. An approach to tackle this issue is to introduce Positional Encoding (PE) of nodes, and inject it into the input layer, like in Transformers. Possible graph PE are Laplacian eigenvectors. In this work, we propose to decouple structural and positional representations to make easy for the network to learn these two essential properties. We introduce a novel generic architecture which we call \texttt{LSPE} (Learnable Structural and Positional Encodings). We investigate several sparse and fully-connected (Transformer-like) GNNs, and observe a performance increase for molecular datasets, from 1.79% up to 64.14% when considering learnable PE for both GNN classes.

****

Disentangling Generalization in Reinforcement Learning
Alex Lewandowski,Dale Schuurmans,Jun Luo

Generalization in Reinforcement Learning (RL) is usually measured according to concepts from supervised learning. Unlike a supervised learning model however, an RL agent must generalize across states, actions and observations from limited reward-based feedback. We propose to measure an RL agent's capacity to generalize by evaluating it in a contextual decision process that combines a tabular environment with observations from a supervised learning dataset. The resulting environment, while simple, necessitates function approximation for state abstraction and provides ground-truth labels for optimal policies and value functions. The ground truth labels provided by our environment enable us to characterize generalization in RL across different axes: state-space, observation-space and action-space. Putting this method to work, we combine the MNIST dataset with various gridworld environments to rigorously evaluate generalization of DQN and QR-DQN in state, observation and action spaces for both online and offline learning. Contrary to previous reports about common regularization methods, we find that dropout does not improve observation generalization. We find, however, that dropout improves action generalization. Our results also corroborate recent findings that QR-DQN is able to generalize to new observations better than DQN in the offline setting. This success does not extend to state generalization, where DQN is able to generalize better than QR-DQN. These findings demonstrate the need for careful consideration of generalization in RL, and we hope that this line of research will continue to shed light on generalization claims in the literature.

****

Code Editing from Few Exemplars by Adaptive Multi-Extent Composition
Peizhao Li,Xuchao Zhang,Ziyu Yao,Wei Cheng,Haifeng Chen,Hongfu Liu

This paper considers the computer source code editing with few exemplars. The editing exemplar, containing the original and modified support code snippets, show cases a certain editorial style and implies the edit intention for a query code snippet. To achieve this, we propose a machine learning approach to adapt the editorial style derived from few exemplars to a query code snippet. Our learning approach combines edit representations extracted from editing exemplars and compositionally generalizes them to the query code snippet editing via multi-extent similarities ensemble. Specifically, we parse the code snippets using language-specific grammar into abstract syntax trees. We apply the similarities measurement in multiple extents from individual nodes to collective tree representations, and ensemble them through a similarity-ranking error estimator. We evaluate the proposed method on two datasets in C\# and Python languages and respectively show 8.0\% and 10.9\% absolute accuracy improvements compared to baselines.

****

Active Learning over Multiple Domains in Natural Language Tasks

Shayne Longpre,Julia Rachel Reisler,Edward Greg Huang,Yi Lu,Andrew Frank,Nikhil Ramesh,Chris DuBois

Studies of active learning traditionally assume the target and source data stem from a single domain. However, in realistic applications, practitioners often require active learning with multiple sources of out-of-distribution data, where it is unclear a priori which data sources will help or hurt the target domain. We survey a wide variety of techniques in active learning (AL), domain shift detection (DS), and multi-domain sampling to examine this challenging setting for question answering and sentiment analysis. We ask (1) what family of methods are effective for this task? And, (2) what properties of selected examples and domains achieve strong results? Among 18 acquisition functions from 4 families of methods, we find H- Divergence methods, and particularly our proposed variant DAL-E, yield effective results, averaging 2-3% improvements over the random baseline. We also show the importance of a diverse allocation of domains, as well as room-for-improvement of existing methods on both domain and example selection. Our findings yield the first comprehensive analysis of both existing and novel methods for practitioners faced with multi-domain active learning for natural language tasks.

****

Provable Federated Adversarial Learning via Min-max Optimization

Xiaoxiao Li,Zhao Song,Jiaming Yang

Federated learning (FL) is a trending training paradigm to utilize decentralized training data. FL allows clients to update model parameters locally for several epochs, then share them to a global model for aggregation. This training paradigm with multi-local step updating before aggregation exposes unique vulnerabilities to adversarial attacks. Adversarial training is a trending method to improve the robustness of neural networks against adversarial perturbations. First, we formulate a \textit{general} form of federated adversarial learning (FAL) that is adapted from adversarial learning in the centralized setting. On the client side of FL training, FAL has an inner loop to optimize an adversarial to generate adversarial samples for adversarial training and an outer loop to update local model parameters. On the server side, FAL aggregates local model updates and broadcast the aggregated model. We design a global training loss to formulate FAL training as a min-max optimization problem. Unlike the convergence analysis in centralized training that relies on the gradient direction, it is significantly harder to analyze the convergence in FAL for two reasons: 1) the complexity of min-max optimization, and 2) model not updating in the gradient direction due to the multi-local updates on the client-side before aggregation. Further, we address the challenges using appropriate gradient approximation and coupling techniques and present the convergence analysis in the over-parameterized regime. Our main result theoretically shows that the minimal value of loss function under this algorithm can converge to ϵ small with chosen learning rate and communication rounds. It is noteworthy that our analysis is feasible for non-IID clients.

****

Less is More: Dimension Reduction Finds On-Manifold Adversarial Examples in Hard-Label Attacks

Washington Garcia,Pin-Yu Chen,Somesh Jha,Hamilton Scott Clouse,Kevin R. B. Butle r

Designing deep networks robust to adversarial examples remains an open problem. Likewise, recent zeroth-order hard-label attacks on image classification models have shown comparable performance to their first-order, gradient-level alternatives. It was recently shown in the gradient-level setting that regular adversarial examples leave the data manifold, while their on-manifold counterparts are in fact generalization errors. In this paper, we argue that query efficiency in the zeroth-order setting is connected to an adversary's traversal through the data manifold. To explain this behavior, we propose an information-theoretic argument based on a noisy manifold distance oracle, which leaks manifold information through the adversary's gradient estimate. Through numerical experiments of manifold-gradient mutual information, we show this behavior acts as a function of the effective problem dimensionality. On high-dimensional real-world datasets and multiple zeroth-order attacks using dimension reduction, we observe the same behavior to produce samples closer to the data manifold. This can result in up to 4x decrease in the manifold distance measure, regardless of the model robustness. Our results suggest that taking the manifold-gradient mutual information into account can thus inform better robust model design in the future, and avoid leakage of the sensitive data manifold information.

****

Exploring and Evaluating Personalized Models for Code Generation

Andrei Zlotchevski,Dawn Drain,Alexey Svyatkovskiy,Colin Clement,Neel Sundaresan,Michele Tufano

Large Transformer models achieved the state-of-the-art status for Natural Language Understanding and are increasingly the baseline architecture for source code generation models. Transformers are usually pre-trained on a large unsupervised corpus, learning token representations and transformations relevant to modeling generally available text, and then fine-tuned on a particular task of interest. While fine-tuning is a tried-and-true method for adapting a model to a new domain, for example question-answering on a given topic or a source code generation model, generalization remains an on-going challenge. Here we explore the ability of various levels of model fine-tuning to improve generalization by personalized fine-tuning. In the context of generating unit tests for Java methods, here we evaluate learning to personalize to a specific project using several methods to personalize transformer models for unit test generation for a specific Java project. We consider three fine-tuning approaches: (i) custom fine-tuning, which allows all the model parameters to be tuned; (ii) lightweight fine-tuning, which freezes most of the model's parameters, allowing a tuning of the token embeddings and softmax layer or the final layer alone; (iii) prefix tuning, which keeps language model parameters frozen, but optimizes a small project-specific prefix vector. Each of these techniques offers a different trade-off in total compute cost and prediction performance, which we evaluate by code and task-specific metrics, training time, and total computational operations. We compare these fine-tuning strategies for code generation and discuss the potential generalization and cost benefits of each in deployment scenarios.

****

Near-Optimal Algorithms for Autonomous Exploration and Multi-Goal Stochastic Shortest Path

Haoyuan Cai,Tengyu Ma,Simon Shaolei Du

We revisit the incremental autonomous exploration problem proposed by Lim and Auer (2012). In this setting, the agent aims to learn a set of near-optimal goal-conditioned policies to reach the L-controllable states: states that are incrementally reachable from an initial state s_0 within L steps in expectation. We introduce three new algorithms with stronger sample complexity bounds than existing ones. Furthermore, we also prove the first lower bound for the autonomous exploration problem. In particular, the lower bound implies that one of our proposed algorithms, Value-Aware Autonomous Exploration, is nearly minimax-optimal when the number of L-controllable states grows polynomially with respect to L. Key in our algorithm design is a connection between autonomous exploration and multi-goal stochastic shortest path, a new problem that naturally generalizes t he classical stochastic shortest path problem. This new problem and its connection to autonomous exploration can be of independent interest.

****

Autoregressive Quantile Flows for Predictive Uncertainty Estimation
Phillip Si,Allan Bishop,Volodymyr Kuleshov
Numerous applications of machine learning involve representing probability distributions over high-dimensional data. We propose autoregressive quantile flows, a flexible class of normalizing flow models trained using a novel objective based on proper scoring rules. Our objective does not require calculating computationally expensive determinants of Jacobians during training and supports new types of neural architectures, such as neural autoregressive flows from which it is easy to sample.

We leverage these models in quantile flow regression, an approach that parameterizes predictive conditional distributions with flows, resulting in improved probabilistic predictions on tasks such as time series forecasting and object detection.

Our novel objective functions and neural flow parameterizations also yield improvements on popular generation and density estimation tasks, and represent a step beyond maximum likelihood learning of flows.

****

Bypassing Logits Bias in Online Class-Incremental Learning with a Generative Framework
Gehui Shen,Shibo Jie,Ziheng Li,Zhi-Hong Deng
Continual learning requires the model to maintain the learned knowledge while learning from a non-i.i.d data stream continually. Due to the single-pass training setting, online continual learning is very challenging, but it is closer to the real-world scenarios where quick adaptation to new data is appealing. In this paper, we focus on online class-incremental learning setting in which new classes emerge over time. Almost all existing methods are replay-based with a softmax classifier. However, the inherent logits bias problem in the softmax classifier is a main cause of catastrophic forgetting while existing solutions are not applicable for online settings. To bypass this problem, we abandon the softmax classifier and propose a novel generative framework based on the feature space. In our framework, a generative classifier which utilizes replay memory is used for inference, and the training objective is a pair-based metric learning loss which is proven theoretically to optimize the feature space in a generative way. In order to improve the ability to learn new data, we further propose a hybrid of generative and discriminative loss to train the model. Extensive experiments on several benchmarks, including newly introduced task-free datasets, show that our method beats a series of state-of-the-art replay-based methods with discriminative classifiers, and reduces catastrophic forgetting consistently with a remarkable margin.

****

Large Language Models Can Be Strong Differentially Private Learners
Xuechen Li,Florian Tramer,Percy Liang,Tatsunori Hashimoto
Differentially Private (DP) learning has seen limited success for building large deep learning models of text, and straightforward attempts at applying Differentially Private Stochastic Gradient Descent (DP-SGD) to NLP tasks have resulted in large performance drops and high computational overhead.

We show that this performance drop can be mitigated with (1) the use of large pretrained language models; (2) non-standard hyperparameters that suit DP optimization; and (3) fine-tuning objectives which are aligned with the pretraining procedure.

With the above, we obtain NLP models that outperform state-of-the-art DP-trained models under the same privacy budget and strong non-private baselines---by directly fine-tuning pretrained models with DP optimization on moderately-sized corpora.

To address the computational challenge of running DP-SGD with large Transformers, we propose a memory saving technique that allows clipping in DP-SGD to run without instantiating per-example gradients for any linear layer in the model.

The technique enables privately training Transformers with almost the same memory cost as non-private training at a modest run-time overhead.
Contrary to conventional wisdom that DP optimization fails at learning high-dimensional models (due to noise that scales with dimension) empirical results reveal that private learning with pretrained language models tends to not suffer from dimension-dependent performance degradation.
Code to reproduce results can be found at https://github.com/lxuechen/private-transformers.

****
G-Mixup: Graph Augmentation for Graph Classification
Xiaotian Han,Zhimeng Jiang,Ninghao Liu,Xia Hu
This work develops \emph{mixup to graph data}. Mixup has shown superiority in improving the generalization and robustness of neural networks by interpolating features and labels of random two samples. Traditionally, Mixup can operate on regular, grid-like, and Euclidean data such as image or tabular data. However, it is challenging to directly adopt Mixup to augment graph data because two graphs typically: 1) have different numbers of nodes; 2) are not readily aligned; and 3) have unique topologies in non-Euclidean space. To this end, we propose $\mathcal{G}$-Mixup to augment graphs for graph classification by interpolating the generator (i.e., graphon) of different classes of graphs. Specifically, we first use graphs within the same class to estimate a graphon. Then, instead of directly manipulating graphs, we interpolate graphons of different classes in the Euclidean space to get mixed graphons, where the synthetic graphs are generated through sampling based on the new graphons.
****
Local Patch AutoAugment with Multi-Agent Collaboration
Shiqi Lin,Tao Yu,Ruoyu Feng,Xin Li,Xin Jin,Zhibo Chen
Data augmentation (DA) plays a critical role in improving the generalization of deep learning models. Recent works on automatically searching for DA policies from data have achieved great success. However, existing automated DA methods generally perform the search at the image level, which limits the exploration of diversity in local regions. In this paper, we propose a more fine-grained automated DA approach, dubbed Patch AutoAugment, to divide an image into a grid of patches and search for the joint optimal augmentation policies for the patches. We formulate it as a multi-agent reinforcement learning (MARL) problem, where each agent learns an augmentation policy for each patch based on its content together with the semantics of the whole image. The agents cooperate with each other to achieve the optimal augmentation effect of the entire image by sharing a team reward. We show the effectiveness of our method on multiple benchmark datasets of image classification and fine-grained image recognition (e.g., CIFAR-10, CIFAR-100, ImageNet, CUB-200-2011, Stanford Cars and FGVC-Aircraft). Extensive experiments demonstrate that our method outperforms the state-of-the-art DA methods while requiring fewer computational resources.
****
Learning Visual-Linguistic Adequacy, Fidelity, and Fluency for Novel Object Captioning
Cheng-Fu Yang,Yao-Hung Hubert Tsai,Wan-Cyuan Fan,Yu-Chiang Frank Wang,Louis-Philippe Morency,Ruslan Salakhutdinov
Novel object captioning (NOC) learns image captioning models for describing objects or visual concepts which are unseen (i.e., novel) in the training captions. Such captioning models need to sufficiently describe such visual data with fluent and natural language expression. In other words, we expect the produced captions being linguistically fluent, containing novel objects of interest, and fitting the visual concept of the image. The above three aspects thus correspond to fluency, fidelity, and adequacy, respectively. However, most novel object captioning models are not explicitly designed to address the aforementioned properties due to the absence of caption annotations. In this paper, we start by providing an insight into the relationship between the above properties and existing visual/language models. Then, we present VLAF2, for learning Visual-Linguistic Adequac

y, Fidelity, and Fluency, which utilizes linguistics observed from captions for describing visual information of images with novel objects. More specifically, we revisit BERT and CLIP, and explain how we leverage the intrinsic language knowledge from such popular models to reward captions with precise and rich visual content associated with novel images. To validate the effectiveness of our framework, we conduct extensive experiments on the nocaps dataset. Our method not only performs favorably against state-of-the-art novel captioning models in all caption evaluation metrics, but also surpasses the SPICE scores of human baseline. We perform quantitative and qualitative analysis to demonstrate how our model generates novel object captions with improved fluency, fidelity, and adequacy. Implementation details and code are available in the supplementary materials.

****

Zero-Shot Self-Supervised Learning for MRI Reconstruction

Burhaneddin Yaman,Seyed Amir Hossein Hosseini,Mehmet Akcakaya

Deep learning (DL) has emerged as a powerful tool for accelerated MRI reconstruction, but often necessitates a database of fully-sampled measurements for training. Recent self-supervised and unsupervised learning approaches enable training without fully-sampled data. However, a database of undersampled measurements may not be available in many scenarios, especially for scans involving contrast or translational acquisitions in development. Moreover, recent studies show that database-trained models may not generalize well when the unseen measurements differ in terms of sampling pattern, acceleration rate, SNR, image contrast, and anatomy. Such challenges necessitate a new methodology to enable subject-specific DL MRI reconstruction without external training datasets, since it is clinically imperative to provide high-quality reconstructions that can be used to identify lesions/disease for $\textit{every individual}$. In this work, we propose a zero-shot self-supervised learning approach to perform subject-specific accelerated DL MRI reconstruction to tackle these issues. The proposed approach partitions the available measurements from a single scan into three disjoint sets. Two of these sets are used to enforce data consistency and define loss during training for self-supervision, while the last set serves to self-validate, establishing an early stopping criterion. In the presence of models pre-trained on a database with different image characteristics, we show that the proposed approach can be combined with transfer learning for faster convergence time and reduced computational complexity.

****

Text Generation with Efficient (Soft) Q-Learning

Han Guo,Bowen Tan,Zhengzhong Liu,Eric Xing,Zhiting Hu

Maximum likelihood estimation (MLE) is the predominant algorithm for training text generation models. This paradigm relies on direct supervision examples, which is not applicable to many emerging applications, such as generating adversarial attacks or generating prompts to control language models. Reinforcement learning (RL) on the other hand offers a more flexible solution by allowing users to plug in arbitrary task metrics as reward. Yet previous RL algorithms for text generation, such as policy gradient (on-policy RL) and Q-learning (off-policy RL), are often notoriously inefficient or unstable to train due to the large sequence space and the sparse reward received only at the end of sequences. In this paper, we introduce a new RL formulation for text generation from the soft Q-learning (SQL) perspective. It enables us to draw from the latest RL advances, such as path consistency learning, to combine the best of on-/off-policy updates, and learn effectively from sparse reward. We apply the approach to a wide range of text generation tasks, including learning from noisy/negative examples, adversarial attacks, and prompt generation. Experiments show our approach consistently outperforms both task-specialized algorithms and the previous RL methods.

****

Variational Component Decoder for Source Extraction from Nonlinear Mixture

Shujie Zhang,Tianyue Zheng,Zhe Chen,Jun Luo,Sinno Pan

In many practical scenarios of signal extraction from a nonlinear mixture, only one (signal) source is intended to be extracted. However, modern methods involving Blind Source Separation are inefficient for this task since they are designed to recover all sources in the mixture. In this paper, we propose supervised Variational Component Decoder (sVCD) as a method dedicated to extracting a single source from nonlinear mixture. sVCD leverages the sequence-to-sequence (Seq2Seq) translation ability of a specially designed neural network to approximate a nonlinear inverse of the mixture process, assisted by priors of the interested source. In order to maintain the robustness in the face of real-life samples, sVCD combines Seq2Seq with variational inference to form a deep generative model, and it is trained by optimizing a variant of variational bound on the data likelihood concerning only the interested source. We demonstrate that sVCD has superior performance on nonlinear source extraction over a state-of-the-art method on diverse datasets, including artificially generated sequences, radio frequency (RF) sensing data, and electroencephalogram (EEG) results.

****

Deep Representations for Time-varying Brain Datasets

Sikun Lin,Shuyun Tang,Ambuj Singh

Finding an appropriate representation of dynamic activities in the brain is crucial for many downstream applications. Due to its highly dynamic nature, temporally averaged fMRI (functional magnetic resonance imaging) cannot capture the whole picture of underlying brain activities, and previous works lack the ability to learn and interpret the latent dynamics in brain architectures. In this paper, we build an efficient graph neural network model that incorporates both region-mapped fMRI sequences and structural connectivities obtained from DWI (diffusion-weighted imaging) as inputs. Through novel sample-level adaptive adjacency matrix learning and multi-resolution inner cluster smoothing, we find good representations of the latent brain dynamics. We also attribute inputs with integrated gradients, which enables us to infer (1) highly involved brain connections and subnetworks for each task (2) keyframes of imaging sequences along the temporal axis, and (3) subnetworks that discriminate between individual subjects. This ability to identify critical subnetworks that characterize brain states across heterogeneous tasks and individuals is of great importance to neuroscience research. Extensive experiments and ablation studies demonstrate our proposed method's superiority and efficiency in spatial-temporal graph signal modeling with insightful interpretations of brain dynamics.

****

Towards Understanding Distributional Reinforcement Learning: Regularization, Optimization, Acceleration and Sinkhorn Algorithm

Ke Sun,Yingnan Zhao,Yi Liu,Enze Shi,Yafei Wang,Aref Sadeghi,Xiaodong Yan,Bei Jiang,Linglong Kong

Distributional reinforcement learning~(RL) is a class of state-of-the-art algorithms that estimate the whole distribution of the total return rather than only its expectation. Despite the remarkable performance of distributional RL, a theoretical understanding of its advantages over expectation-based RL remains elusive. In this paper, we interpret distributional RL as entropy-regularized maximum likelihood estimation in the \textit{neural Z-fitted iteration} framework and establish the connection of the resulting risk-aware regularization with maximum entropy RL. In addition, We shed light on the stability-promoting distributional loss with desirable smoothness properties in distributional RL, which can yield stable optimization and guaranteed generalization. We also analyze the acceleration behavior while optimizing distributional RL algorithms and show that an appropriate approximation to the true target distribution can speed up the convergence. From the perspective of representation, we find that distributional RL encourages state representation from the same action class classified by the policy in tighter clusters. Finally, we propose a class of \textit{Sinkhorn distributional RL} algorithm that interpolates between the Wasserstein distance and maximum mean discrepancy~(MMD). Experiments on a suite of Atari games reveal the competitive performance of our algorithm relative to existing state-of-the-art distributional RL algorithms.

****

GeoDiff: A Geometric Diffusion Model for Molecular Conformation Generation

Minkai Xu,Lantao Yu,Yang Song,Chence Shi,Stefano Ermon,Jian Tang

Predicting molecular conformations from molecular graphs is a fundamental problem in cheminformatics and drug discovery. Recently, significant progress has been achieved with machine learning approaches, especially with deep generative models. Inspired by the diffusion process in classical non-equilibrium thermodynamics where heated particles will diffuse from original states to a noise distribution, in this paper, we propose a novel generative model named GeoDiff for molecular conformation prediction. GeoDiff treats each atom as a particle and learns to directly reverse the diffusion process (i.e., transforming from a noise distribution to stable conformations) as a Markov chain. Modeling such a generation process is however very challenging as the likelihood of conformations should be roto-translational invariant. We theoretically show that Markov chains evolving with equivariant Markov kernels can induce an invariant distribution by design, and further propose building blocks for the Markov kernels to preserve the desirable equivariance property. The whole framework can be efficiently trained in an end-to-end fashion by optimizing a weighted variational lower bound to the (conditional) likelihood. Experiments on multiple benchmarks show that GeoDiff is superior or comparable to existing state-of-the-art approaches, especially on large molecules.

****

Policy Smoothing for Provably Robust Reinforcement Learning
Aounon Kumar,Alexander Levine,Soheil Feizi

The study of provable adversarial robustness for deep neural networks (DNNs) has mainly focused on $\textit{static}$ supervised learning tasks such as image classification. However, DNNs have been used extensively in real-world $\textit{adaptive}$ tasks such as reinforcement learning (RL), making such systems vulnerable to adversarial attacks as well. Prior works in provable robustness in RL seek to certify the behaviour of the victim policy at every time-step against a non-adaptive adversary using methods developed for the static setting. But in the real world, an RL adversary can infer the defense strategy used by the victim agent by observing the states, actions, etc. from previous time-steps and adapt itself to produce stronger attacks in future steps (e.g., by focusing more on states critical to the agent's performance). We present an efficient procedure, designed specifically to defend against an adaptive RL adversary, that can directly certify the total reward without requiring the policy to be robust at each time-step. Focusing on randomized smoothing based defenses, our main theoretical contribution is to prove an $\textit{adaptive version}$ of the Neyman-Pearson Lemma -- a key lemma for smoothing-based certificates -- where the adversarial perturbation at a particular time can be a stochastic function of current and previous observations and states as well as previous actions. Building on this result, we propose $\textit{policy smoothing}$ where the agent adds a Gaussian noise to its observation at each time-step before passing it through the policy function. Our robustness certificates guarantee that the final total reward obtained by policy smoothing remains above a certain threshold, even though the actions at intermediate time-steps may change under the attack. We show that our certificates are $\textit{tight}$ by constructing a worst-case scenario that achieves the bounds derived in our analysis. Our experiments on various environments like Cartpole, Pong, Freeway and Mountain Car show that our method can yield meaningful robustness guarantees in practice.

****

The Close Relationship Between Contrastive Learning and Meta-Learning
Renkun Ni,Manli Shu,Hossein Souri,Micah Goldblum,Tom Goldstein

Contrastive learning has recently taken off as a paradigm for learning from unlabeled data. In this paper, we discuss the close relationship between contrastive learning and meta-learning under a certain task distribution. We complement this observation by showing that established meta-learning methods, such as Prototypical Networks, achieve comparable performance to SimCLR when paired with this task distribution. This relationship can be leveraged by taking established techniques from meta-learning, such as task-based data augmentation, and showing that they benefit contrastive learning as well. These tricks also benefit state-of-t he-art self-supervised learners without using negative pairs such as BYOL, which achieves 94.6\% accuracy on CIFAR-10 using a self-supervised ResNet-18 feature extractor trained with our meta-learning tricks. We conclude that existing advances designed for contrastive learning or meta-learning can be exploited to benefit the other, and it is better for contrastive learning researchers to take lessons from the meta-learning literature (and vice-versa) than to reinvent the wheel.

****

Efficient Out-of-Distribution Detection via CVAE data Generation

Mengyu Wang,Yijia Shao,Haowei Lin,Wenpeng Hu,Bing Liu

Recently, contrastive loss with data augmentation and pseudo class creation has been shown to produce markedly better results for out-of-distribution (OOD) detection than previous methods. However, a major shortcoming of this approach is that it is extremely slow due to significant increase in the data size and the number of classes and the quadratic complexity of pairwise similarity computation. This paper proposes a novel and simple method that can build an effective data generator using Conditional Variational Auto-Encoder (CVAE) to generate pseudo OOD samples. Based on the generated pseudo OOD data, a flexible and efficient OOD detection method is proposed through fine-tuning, which achieves results comparable to the state-of-the-art OOD detection techniques, but the execution speed is at least 10 times faster. Also importantly, the proposed approach is in fact a general framework that can be applied to many existing OOD methods and improve them via the proposed fine-tuning. We have combined it with the best baseline OOD models in our experiments to produce new state-of-the-art results.

****

FedDrop: Trajectory-weighted Dropout for Efficient Federated Learning

Dongping Liao,Xitong Gao,Yiren Zhao,Hao Dai,Li Li,Kafeng Wang,Kejiang Ye,Yang Wang,Cheng-zhong Xu

Federated learning (FL) enables edge clients to train collaboratively while preserving individual's data privacy. As clients do not inherently share identical data distributions, they may disagree in the direction of parameter updates, resulting in high compute and communication costs in comparison to centralized learning. Recent advances in FL focus on reducing data transmission during training; yet they neglected the increase of computational cost that dwarfs the merit of reduced communication. To this end, we propose FedDrop, which introduces channel-wise weighted dropout layers between convolutions to accelerate training while minimizing their impact on convergence. Empirical results show that FedDrop can drastically reduce the amount of FLOPs required for training with a small increase in communication, and push the Pareto frontier of communication/computation trade-off further than competing FL algorithms.

****

SeqPATE: Differentially Private Text Generation via Knowledge Distillation

Zhiliang Tian,Yingxiu Zhao,Ziyue Huang,Yu-Xiang Wang,Nevin Zhang,He He

Protecting the privacy of user data is crucial when training neural text generation models, which may leak sensitive user information during generation. Differentially private (DP) learning algorithms provide guarantees on identifying the existence of a training sample from model outputs. PATE is a DP learning algorithm that fits the large model well, such as GPT. In this paper, we propose SeqPATE that adapts PATE to text generation while satisfying DP. There are two key challenges in adapting PATE to text generation: (i) obtaining sequence-level supervision for text generation, and (ii) reducing noise required to protect privacy given the large output space (i.e. vocabulary size). For (i), we generate pseudo input and reduce the sequence generation problem to the next word prediction. For (ii), we reduce the output space with top-k and top-p selection strategy that dynamically filters the candidate words; and we refine the teacher aggregation mechanism of PATE to avoid the low agreement rates due to voting over the large output space. To limit the privacy loss, we design an efficient knowledge distillation to reduce the time of distilling from the private data. We apply SeqPATE to a simple text generation task (sentence completion) and achieves 39\% and 28\% gains in Bleu4 on two datasets.

****

Multi-Resolution Continuous Normalizing Flows

Vikram Voleti,Chris Finlay,Adam M Oberman,Christopher Pal

Recent work has shown that Neural Ordinary Differential Equations (ODEs) can serve as generative models of images using the perspective of Continuous Normalizing Flows (CNFs). Such models offer exact likelihood calculation, and invertible generation/density estimation. In this work we introduce a Multi-Resolution variant of such models (MRCNF), by characterizing the conditional distribution over the additional information required to generate a fine image that is consistent with the coarse image. We introduce a transformation between resolutions that allows for no change in the log likelihood. We show that this approach yields comparable likelihood values for various image datasets, using orders of magnitude fewer parameters than the prior methods, in significantly less training time, using only one GPU.

****

Towards Understanding Generalization via Decomposing Excess Risk Dynamics

Jiaye Teng,Jianhao Ma,Yang Yuan

Generalization is one of the fundamental issues in machine learning. However, traditional techniques like uniform convergence may be unable to explain generalization under overparameterization \citep{nagarajan2019uniform}. As alternative approaches, techniques based on stability analyze the training dynamics and derive algorithm-dependent generalization bounds. Unfortunately, the stability-based bounds are still far from explaining the surprising generalization in deep learning since neural networks usually suffer from unsatisfactory stability. This paper proposes a novel decomposition framework to improve the stability-based bounds via a more fine-grained analysis of the signal and noise, inspired by the observation that neural networks converge relatively slowly when fitting noise (which indicates better stability). Concretely, we decompose the excess risk dynamics and apply the stability-based bound only on the noise component. The decomposition framework performs well in both linear regimes (overparameterized linear regression) and non-linear regimes (diagonal matrix recovery). Experiments on neural networks verify the utility of the decomposition framework.

****

On the Importance of Firth Bias Reduction in Few-Shot Classification

Saba Ghaffari,Ehsan Saleh,David Forsyth,Yu-Xiong Wang

Learning accurate classifiers for novel categories from very few examples, known as few-shot image classification, is a challenging task in statistical machine learning and computer vision. The performance in few-shot classification suffers from the bias in the estimation of classifier parameters; however, an effective underlying bias reduction technique that could alleviate this issue in training few-shot classifiers has been overlooked. In this work, we demonstrate the effectiveness of Firth bias reduction in few-shot classification. Theoretically, Firth bias reduction removes the $O(N^{-1})$ first order term from the small-sample bias of the Maximum Likelihood Estimator. Here we show that the general Firth bias reduction technique simplifies to encouraging uniform class assignment probabilities for multinomial logistic classification, and almost has the same effect in cosine classifiers. We derive an easy-to-implement optimization objective for Firth penalized multinomial logistic and cosine classifiers, which is equivalent to penalizing the cross-entropy loss with a KL-divergence between the predictions and the uniform label distribution. Then, we empirically evaluate that it is consistently effective across the board for few-shot image classification, regardless of (1) the feature representations from different backbones, (2) the number of samples per class, and (3) the number of classes. Furthermore, we demonstrate the effectiveness of Firth bias reduction on cross-domain and imbalanced data settings. Our implementation is available at https://github.com/ehsansaleh/firth_bias_reduction.

****

Graph Auto-Encoder via Neighborhood Wasserstein Reconstruction

Mingyue Tang,Pan Li,Carl Yang

Graph neural networks (GNNs) have drawn significant research attention recently, mostly under the setting of semi-supervised learning. When task-agnostic representations are preferred or supervision is simply unavailable, the auto-encoder framework comes in handy with a natural graph reconstruction objective for unsupervised GNN training. However, existing graph auto-encoders are designed to reconstruct the direct links, so GNNs trained in this way are only optimized towards proximity-oriented graph mining tasks, and will fall short when the topological structures matter. In this work, we revisit the graph encoding process of GNNs which essentially learns to encode the neighborhood information of each node into an embedding vector, and propose a novel graph decoder to reconstruct the entire neighborhood information regarding both proximity and structure via Neighborhood Wasserstein Reconstruction (NWR). Specifically, from the GNN embedding of each node, NWR jointly predicts its node degree and neighbor feature distribution, where the distribution prediction adopts an optimal-transport loss based on the Wasserstein distance. Extensive experiments on both synthetic and real-world network datasets show that the unsupervised node representations learned with NWR have much more advantageous in structure-oriented graph mining tasks, while also achieving competitive performance in proximity-oriented ones.

****

OVD-Explorer: A General Information-theoretic Exploration Approach for Reinforcement Learning

Jinyi Liu,Zhi Wang,YAN ZHENG,Jianye HAO,Junjie Ye,Chenjia Bai,Pengyi Li

Many exploration strategies are built upon the optimism in the face of the uncertainty (OFU) principle for reinforcement learning. However, without considering the aleatoric uncertainty, existing methods may over-explore the state-action pairs with large randomness and hence are non-robust. In this paper, we explicitly capture the aleatoric uncertainty from a distributional perspective and propose an information-theoretic exploration method named Optimistic Value Distribution Explorer (OVD-Explorer). OVD-Explorer follows the OFU principle, but more importantly, it avoids exploring the areas with high aleatoric uncertainty through maximizing the mutual information between policy and the upper bounds of policy's returns. Furthermore, to make OVD-Explorer tractable for continuous RL, we derive a closed form solution, and integrate it with SAC, which, to our knowledge, for the first time alleviates the negative impact on exploration caused by aleatoric uncertainty for continuous RL. Empirical evaluations on the commonly used Mujoco benchmark and a novel GridChaos task demonstrate that OVD-Explorer can alleviate over-exploration and outperform state-of-the-art methods.

****

Finite-Time Convergence and Sample Complexity of Multi-Agent Actor-Critic Reinforcement Learning with Average Reward

FNU Hairi,Jia Liu,Songtao Lu

In this paper, we establish the first finite-time convergence result of the actor-critic algorithm for fully decentralized multi-agent reinforcement learning (MARL) problems with average reward.

In this problem, a set of N agents work cooperatively to maximize the global average reward through interacting with their neighbors over a communication network.

We consider a practical MARL setting, where the rewards and actions of each agent are only known to itself, and the knowledge of joint actions of the agents is not assumed.

Toward this end, we propose a mini-batch Markovian sampled fully decentralized actor-critic algorithm and analyze its finite-time convergence and sample complexity.

We show that the sample complexity of this algorithm is $\mathcal{O}(N^{2}/\epsilon^{2}\log(N/\epsilon))$.

Interestingly, this sample complexity bound matches that of the state-of-the-art single-agent actor-critic algorithms for reinforcement learning.

****

Non-convex Optimization for Learning a Fair Predictor under Equalized Loss Fairness Constraint

Mohammad Mahdi Khalili,Xueru Zhang,Mahed Abroshan,Iman Vakilinia

Supervised learning models have been increasingly used in various domains such as lending, college admission, natural language processing, face recognition, etc. These models may inherit pre-existing biases from training datasets and exhibit discrimination against protected social groups. Various fairness notions have been introduced to address fairness issues. In general, finding a fair predictor leads to a constrained optimization problem, and depending on the fairness notion, it may be non-convex. In this work, we focus on Equalized Loss ($\textsf{EL}$), a fairness notion that requires the prediction error/loss to be equalized across different demographic groups. Imposing this constraint to the learning process leads to a non-convex optimization problem even if the loss function is convex. We introduce algorithms that can leverage off-the-shelf convex programming tools and efficiently find the $\textit{global}$ optimum of this non-convex problem. In particular, we first propose the $\mathtt{ELminimizer}$ algorithm, which finds the optimal $\textsf{EL}$ fair predictor by reducing the non-convex optimization problem to a sequence of convex constrained optimizations. We then propose a simple algorithm that is computationally more efficient compared to $\mathtt{ELminimizer}$ and finds a sub-optimal $\textsf{EL}$ fair predictor using $\textit{unconstrained}$ convex programming tools. Experiments on real-world data show the effectiveness of our algorithms.

****

FairCal: Fairness Calibration for Face Verification

Tiago Salvador,Stephanie Cairns,Vikram Voleti,Noah Marshall,Adam M Oberman

Despite being widely used, face recognition models suffer from bias: the probability of a false positive (incorrect face match) strongly depends on sensitive attributes such as the ethnicity of the face. As a result, these models can disproportionately and negatively impact minority groups, particularly when used by law enforcement. The majority of bias reduction methods have several drawbacks: they use an end-to-end retraining approach, may not be feasible due to privacy issues, and often reduce accuracy. An alternative approach is post-processing methods that build fairer decision classifiers using the features of pre-trained models, thus avoiding the cost of retraining. However, they still have drawbacks: they reduce accuracy (AGENDA, FTC), or require retuning for different false positive rates (FSN). In this work, we introduce the Fairness Calibration (FairCal) method, a post-training approach that simultaneously: (i) increases model accuracy (improving the state-of-the-art), (ii) produces fairly-calibrated probabilities, (iii) significantly reduces the gap in the false positive rates, (iv) does not require knowledge of the sensitive attribute, and (v) does not require retraining, training an additional model or retuning. We apply it to the task of Face Verification, and obtain state-of-the-art results with all the above advantages.

****

Cross-Lingual Transfer with Class-Weighted Language-Invariant Representations

Ruicheng Xian,Heng Ji,Han Zhao

Recent advances in neural modeling have produced deep multilingual language models capable of extracting cross-lingual knowledge from non-parallel texts and enabling zero-shot downstream transfer. While their success is often attributed to shared representations, quantitative analyses are limited. Towards a better understanding, through empirical analyses, we show that the invariance of feature representations across languages—an effect of shared representations—strongly correlates with transfer performance. We also observe that distributional shifts in class priors between source and target language task data negatively affect performance, a largely overlooked issue that could cause negative transfer with existing unsupervised approaches. Based on these findings, we propose and evaluate a method for unsupervised transfer, called importance-weighted domain alignment (IWDA), that performs representation alignment with prior shift estimation and correction using unlabeled target language task data. Experiments demonstrate its superiority under large prior shifts, and show further performance gains when combined with existing semi-supervised learning techniques.

****

ComPhy: Compositional Physical Reasoning of Objects and Events from Videos

Zhenfang Chen,Kexin Yi,Yunzhu Li,Mingyu Ding,Antonio Torralba,Joshua B. Tenenbau m,Chuang Gan

Objects' motions in nature are governed by complex interactions and their properties. While some properties, such as shape and material, can be identified via the object's visual appearances, others like mass and electric charge are not directly visible. The compositionality between the visible and hidden properties poses unique challenges for AI models to reason from the physical world, whereas humans can effortlessly infer them with limited observations. Existing studies on video reasoning mainly focus on visually observable elements such as object appearance, movement, and contact interaction. In this paper, we take an initial step to highlight the importance of inferring the hidden physical properties not directly observable from visual appearances, by introducing the Compositional Physical Reasoning (ComPhy) dataset. For a given set of objects, ComPhy includes few videos of them moving and interacting under different initial conditions. The model is evaluated based on its capability to unravel the compositional hidden properties, such as mass and charge, and use this knowledge to answer a set of questions posted on one of the videos. Evaluation results of several state-of-the-art video reasoning models on ComPhy show unsatisfactory performance as they fail to capture these hidden properties. We further propose an oracle neural-symbolic framework named Compositional Physics Learner (CPL), combining visual perception, physical property learning, dynamic prediction, and symbolic execution into a unified framework. CPL can effectively identify objects' physical properties from their interactions and predict their dynamics to answer questions.

****

An Information Fusion Approach to Learning with Instance-Dependent Label Noise

Zhimeng Jiang,Kaixiong Zhou,Zirui Liu,Li Li,Rui Chen,Soo-Hyun Choi,Xia Hu

Instance-dependent label noise (IDN) widely exists in real-world datasets and usually misleads the training of deep neural networks. Noise transition matrix (NTM) (i.e., the probability that clean labels flip into noisy labels) is used to characterize the label noise and can be adopted to bridge the gap between clean and noisy underlying data distributions. However, most instances are long-tail, i.e., the number of occurrences of each instance is usually limited, which leads to the gap between the underlying distribution and the empirical distribution. Therefore, the genuine problem caused by IDN is \emph{empirical}, instead of underlying, \emph{data distribution mismatch} during training. To directly tackle the empirical distribution mismatch problem, we propose \emph{posterior transition matrix} (PTM) to posteriorly model label noise given limited observed noisy labels, which achieves \emph{statistically consistent classifiers}. Note that even if an instance is corrupted by the same NTM, the intrinsic randomness incurs different noisy labels, and thus requires different correction methods. Motivated by this observation, we propose an \textbf{I}nformation \textbf{F}usion (IF) approach to fine-tune the NTM based on the estimated PTM. Specifically, we adopt the noisy labels and model predicted probabilities to estimate the PTM and then correct the NTM in \emph{forward propagation}. Empirical evaluations on synthetic and real-world datasets demonstrate that our method is superior to the state-of-the-art approaches, and achieves more stable training for instance-dependent label noise.

****

Deep Reinforcement Learning for Equal Risk Option Pricing and Hedging under Dynamic Expectile Risk Measures

Saeed Marzban,Erick Delage,Jonathan Li

Recently equal risk pricing, a framework for fair derivative pricing, was extended to consider coherent risk measures. However, all current implementations either employ a static risk measure or are based on traditional dynamic programming solution schemes that are impracticable in realistic settings: when the number of underlying assets is large or only historical trajectories are available. This paper extends for the first time the deep deterministic policy gradient algorithm to the problem of solving a risk averse Markov decision process that models risk using a time consistent dynamic expectile risk measure. Our numerical experiments, which involve both a simple vanilla option and a more exotic basket option, confirm that the new ACRL algorithm can produce high quality hedging strateg ies that produce accurate prices in simple settings, and outperform the strategies produced using static risk measures when the risk is evaluated at later points of time.

****

On Redundancy and Diversity in Cell-based Neural Architecture Search

Xingchen Wan,Binxin Ru,Pedro M Esperança,Zhenguo Li

Searching for the architecture cells is a dominant paradigm in NAS. However, little attention has been devoted to the analysis of the cell-based search spaces even though it is highly important for the continual development of NAS.

In this work, we conduct an empirical post-hoc analysis of architectures from the popular cell-based search spaces and find that the existing search spaces contain a high degree of redundancy: the architecture performance is less sensitive to changes at large parts of the cells, and universally adopted design rules, like the explicit search for a reduction cell, significantly increase the complexities but have very limited impact on the performance.

Across architectures found by a diverse set of search strategies, we consistently find that the parts of the cells that do matter for architecture performance often follow similar and simple patterns. By constraining cells to include these patterns, randomly sampled architectures can match or even outperform the state of the art.

These findings cast doubts into our ability to discover truly novel architectures in the existing cell-based search spaces and, inspire our suggestions for improvement to guide future NAS research.

Code is available at https://github.com/xingchenwan/cell-based-NAS-analysis.

****

Deep Learning without Shortcuts: Shaping the Kernel with Tailored Rectifiers

Guodong Zhang,Aleksandar Botev,James Martens

Training very deep neural networks is still an extremely challenging task. The common solution is to use shortcut connections and normalization layers, which are both crucial ingredients in the popular ResNet architecture. However, there is strong evidence to suggest that ResNets behave more like ensembles of shallower networks than truly deep ones. Recently, it was shown that deep vanilla networks (i.e.~networks without normalization layers or shortcut connections) can be trained as fast as ResNets by applying certain transformations to their activation functions. However, this method (called Deep Kernel Shaping) isn't fully compatible with ReLUs, and produces networks that overfit significantly more than ResNets on ImageNet. In this work, we rectify this situation by developing a new type of transformation that is fully compatible with a variant of ReLUs -- Leaky ReLUs. We show in experiments that our method, which introduces negligible extra computational cost, achieves validation accuracies with deep vanilla networks that are competitive with ResNets (of the same width/depth), and significantly higher than those obtained with the Edge of Chaos (EOC) method. And unlike with EOC, the validation accuracies we obtain do not get worse with depth.

****

Towards Understanding the Data Dependency of Mixup-style Training

Muthu Chidambaram,Xiang Wang,Yuzheng Hu,Chenwei Wu,Rong Ge

In the Mixup training paradigm, a model is trained using convex combinations of data points and their associated labels. Despite seeing very few true data points during training, models trained using Mixup seem to still minimize the original empirical risk and exhibit better generalization and robustness on various tasks when compared to standard training. In this paper, we investigate how these benefits of Mixup training rely on properties of the data in the context of classification. For minimizing the original empirical risk, we compute a closed form for the Mixup-optimal classification, which allows us to construct a simple dataset on which minimizing the Mixup loss leads to learning a classifier that does not minimize the empirical loss on the data. On the other hand, we also give sufficient conditions for Mixup training to also minimize the original empirical risk. For generalization, we characterize the margin of a Mixup classifier, and use this to understand why the decision boundary of a Mixup classifier can adapt better to the full structure of the training data when compared to standard train ing. In contrast, we also show that, for a large class of linear models and linearly separable datasets, Mixup training leads to learning the same classifier as standard training.

****

Variational autoencoders in the presence of low-dimensional data: landscape and implicit bias

Frederic Koehler,Viraj Mehta,Chenghui Zhou,Andrej Risteski

Variational Autoencoders (VAEs) are one of the most commonly used generative models, particularly for image data. A prominent difficulty in training VAEs is data that is supported on a lower dimensional manifold. Recent work by Dai and Wipf (2020) proposes a two-stage training algorithm for VAEs, based on a conjecture that in standard VAE training the generator will converge to a solution with 0 variance which is correctly supported on the ground truth manifold. They gave partial support for this conjecture by showing that some optima of the VAE loss do satisfy this property, but did not analyze the training dynamics. In this paper, we show that for linear encoders/decoders, the conjecture is true—that is the VAE training does recover a generator with support equal to the ground truth manifold—and does so due to an implicit bias of gradient descent rather than merely the VAE loss itself. In the nonlinear case, we show that VAE training frequently learns a higher-dimensional manifold which is a superset of the ground truth manifold.

****

Learning Stochastic Shortest Path with Linear Function Approximation

Yifei Min,Jiafan He,Tianhao Wang,Quanquan Gu

We study the stochastic shortest path (SSP) problem in reinforcement learning with linear function approximation, where the transition kernel is represented as a linear mixture of unknown models. We call this class of SSP problems as linear mixture SSP. We propose a novel algorithm for learning the linear mixture SSP, which can attain a $\tilde O(dB_{\star}^{1.5}\sqrt{K/c_{\min}})$ regret. Here K is the number of episodes, d is the dimension of the feature mapping in the mixture model, $B_{\star}$ bounds the expected cumulative cost of the optimal policy, and $c_{\min}>0$ is the lower bound of the cost function. Our algorithm also applies to the case when $c_{\min} = 0$, where a $\tilde O(K^{2/3})$ regret is guaranteed. To the best of our knowledge, this is the first algorithm with a sublinear regret guarantee for learning linear mixture SSP. In complement to the regret upper bounds, we also prove a lower bound of $\Omega(dB_{\star} \sqrt{K})$, which nearly matches our upper bound.

****

No Parameters Left Behind: Sensitivity Guided Adaptive Learning Rate for Training Large Transformer Models

Chen Liang,Haoming Jiang,Simiao Zuo,Pengcheng He,Xiaodong Liu,Jianfeng Gao,Weizhu Chen,Tuo Zhao

Recent research has shown the existence of significant redundancy in large Transformer models. One can prune the redundant parameters without significantly sacrificing the generalization performance. However, we question whether the redundant parameters could have contributed more if they were properly trained. To answer this question, we propose a novel training strategy that encourages all parameters to be trained sufficiently. Specifically, we adaptively adjust the learning rate for each parameter according to its sensitivity, a robust gradient-based measure reflecting this parameter's contribution to the model performance. A parameter with low sensitivity is redundant, and we improve its fitting by increasing its learning rate. In contrast, a parameter with high sensitivity is well-trained, and we regularize it by decreasing its learning rate to prevent further overfitting. We conduct extensive experiments on natural language understanding, neural machine translation, and image classification to demonstrate the effectiveness of the proposed schedule. Analysis shows that the proposed schedule indeed reduces the redundancy and improves generalization performance.

****

Data Scaling Laws in NMT: The Effect of Noise and Architecture

Yamini Bansal,Behrooz Ghorbani,Ankush Garg,Biao Zhang,Colin Cherry,Maxim Krikun, Behnam Neyshabur,Orhan Firat

In this work, we empirically study the data scaling properties of neural machine translation (NMT). We first establish that the test loss of encoder-decoder transformer models scales as a power law in the number of training samples, with a dependence on the model size. We then systematically vary various aspects of the training setup to understand how they impact the data scaling laws. In particular, we change the (1) Architecture and task setup, to a Transformer-LSTM Hybrid as well as a Decoder-only transformer with language modeling loss (2) Noise level in the training distribution, starting with noisy data with filtering applied as well as clean data corrupted with synthetic iid noise. In all the above cases, we find that the data scaling exponents are minimally impacted, suggesting that marginally worse architectures or training data quality can be compensated for by adding more data. Lastly, we find that changing the training distribution to use back-translated data instead of parallel data, can impact the scaling exponent.

****

Learning to Give Checkable Answers with Prover-Verifier Games

Cem Anil,Guodong Zhang,Yuhuai Wu,Roger Baker Grosse

Our ability to know when to trust the decisions made by machine learning systems has not kept up with the staggering improvements in their performance, limiting their applicability in high-stakes applications. We propose Prover-Verifier Games (PVGs), a game-theoretic framework to encourage neural networks to solve decision problems in a verifiable manner. The PVG consists of two learners with competing objectives: a trusted verifier network tries to choose the correct answer, and a more powerful but untrusted prover network attempts to persuade the verifier of a particular answer, regardless of its correctness. The goal is for a reliable justification protocol to emerge from this game. We analyze several variants of the basic framework, including both simultaneous and sequential games, and narrow the space down to a subset of games which provably have the desired equilibria. We then develop practical instantiations of the PVG for several algorithmic tasks, and show that in practice, the verifier is able to receive useful and reliable information from an untrusted prover. Importantly, the protocol still works even when the verifier is frozen and the prover's message is directly optimized to convince the verifier.

****

SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations

Chenlin Meng,Yutong He,Yang Song,Jiaming Song,Jiajun Wu,Jun-Yan Zhu,Stefano Ermon

Guided image synthesis enables everyday users to create and edit photo-realistic images with minimum effort. The key challenge is balancing faithfulness to the user inputs (e.g., hand-drawn colored strokes) and realism of the synthesized images. Existing GAN-based methods attempt to achieve such balance using either conditional GANs or GAN inversions, which are challenging and often require additional training data or loss functions for individual applications. To address these issues, we introduce a new image synthesis and editing method, Stochastic Differential Editing (SDEdit), based on a diffusion model generative prior, which synthesizes realistic images by iteratively denoising through a stochastic differential equation (SDE). Given an input image with user guide in a form of manipulating RGB pixels, SDEdit first adds noise to the input, then subsequently denoises the resulting image through the SDE prior to increase its realism. SDEdit does not require task-specific training or inversions and can naturally achieve the balance between realism and faithfulness. SDEdit outperforms state-of-the-art GAN-based methods by up to 98.09% on realism and 91.72% on overall satisfaction scores, according to a human perception study, on multiple tasks, including stroke-based image synthesis and editing as well as image compositing.

****

Max-Affine Spline Insights Into Deep Network Pruning

Randall Balestriero,Haoran You,Zhihan Lu,Yutong Kou,Huihong Shi,Yingyan Lin,Richard Baraniuk

State-of-the-art (SOTA) approaches to deep network (DN) training overparametrize the model and then prune a posteriori to obtain a "winning ticket'' subnetwork that can be trained from scratch to achieve high accuracy. To date, the literature has remained largely empirical and hence provides little insights into how pruning affects a DN's decision boundary and no guidance regarding how to design a principled pruning technique. Using a recently developed spline interpretation of DNs, we develop new theory and visualization tools that provide new insights into how pruning DN nodes affects the decision boundary. We discover that a DN's spline mappings exhibit an early-bird (EB) phenomenon whereby the spline's partition converges at early training stages, bridging the recently developed max-affine spline theory and lottery ticket hypothesis of DNs. We leverage this new insight to develop a principled and efficient pruning strategy that focuses on a tiny fraction of DN nodes whose corresponding spline partition regions actually contribute to the final decision boundary. Extensive experiments on four networks and three datasets validate that our new spline-based DN pruning approach reduces training FLOPs by up to 3.5x while achieving similar or even better accuracy than state-of-the-art methods. All the codes will be released publicly upon acceptance.

****

Learning Neural Acoustic Fields

Andrew Luo,Yilun Du,Michael J. Tarr,Joshua B. Tenenbaum,Antonio Torralba,Chuang Gan

Our sensory perception of the world is rich and multimodal. When we walk into a cathedral, acoustics as much as appearance inform us of the sanctuary's wide open space. Similarly, when we drop a wineglass, the sound immediately informs us as to whether it has shattered or not. In this vein, 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 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 the acoustic properties of the scene as a linear time-invariant system, NAFs continuously map all emitter and listener location pairs to an impulse response function that can then be applied to new sounds. We demonstrate that NAFs capture environment reverberations of a scene with high fidelity and can predict sound propagation for novel locations. Leveraging the scene structure learned by NAFs, we also demonstrate improved cross-modal generation of novel views of the scene given sparse visual views. Finally, the continuous nature of NAFs enables potential downstream applications such as sound source localization.

****

Post hoc Explanations may be Ineffective for Detecting Unknown Spurious Correlation

Julius Adebayo,Michael Muelly,Harold Abelson,Been Kim

We investigate whether three types of post hoc model explanations–feature attribution, concept activation, and training point ranking–are effective for detecting a model’s reliance on spurious signals in the training data. Specifically, we consider the scenario where the spurious signal to be detected is unknown, at test-time, to the user of the explanation method. We design an empirical methodology that uses semi-synthetic datasets along with pre-specified spurious artifacts to obtain models that verifiably rely on these spurious training signals. We then provide a suite of metrics that assess an explanation method’s reliability for spurious signal detection under various conditions. We find that the post hoc explanation methods tested are ineffective when the spurious artifact is unknown at test-time especially for non-visible artifacts like a background blur. Further, we find that feature attribution methods are susceptible to erroneously indicating dependence on spurious signals even when the model being explained does not rely on spurious artifacts. This finding casts doubt on the utility of these approaches, in the hands of a practitioner, for detecting a model’s reliance on spurious signals.

****

iPrune: A Magnitude Based Unstructured Pruning Method for Efficient Binary Networks in Hardware
Adithya Venkateswaran,Jean-Pierre David
Modern image recognition models span millions of parameters occupying several megabytes and sometimes gigabytes of space, making it difficult to run on resource constrained edge hardware. Binary Neural Networks address this problem by reducing the memory requirements (one single bit per weight and/or activation). The computation requirement and power consumption are also reduced accordingly. Nevertheless, each neuron in such networks has a large number of inputs, making it difficult to implement them efficiently in binary hardware accelerators, especially LUT-based approaches.

In this work, we present a pruning algorithm and associated results on convolutional and dense layers from aforementioned binary networks. We reduce the computation by 4-70x and the memory by 190-2200x with less than 2% loss of accuracy on MNIST and less than 3% loss of accuracy on CIFAR-10 compared to full precision, fully connected equivalents. Compared to very recent work on pruning for binary networks, we still have a gain of 1% on the precision and up to 30% reduction in memory (526KiB vs 750KiB).
****
Benign Overfitting in Adversarially Robust Linear Classification
Jinghui Chen,Yuan Cao,Quanquan Gu
``Benign overfitting'', where classifiers memorize noisy training data yet still achieve a good generalization performance, has drawn great attention in the machine learning community. To explain this surprising phenomenon, a series of works have provided theoretical justification in over-parameterized linear regression, classification, and kernel methods. However, it is not clear if benign overfitting still occurs in the presence of adversarial examples, i.e., examples with tiny and intentional perturbations to fool the classifiers. In this paper, we show that benign overfitting indeed occurs in adversarial training, a principled approach to defend against adversarial examples. In detail, we prove the risk bounds of the adversarially trained linear classifier on the mixture of sub-Gaussian data under ℓ_p adversarial perturbations. Our result suggests that under moderate perturbations, adversarially trained linear classifiers can achieve the near-optimal standard and adversarial risks, despite overfitting the noisy training data. Numerical experiments validate our theoretical findings.
****
Generalizing Few-Shot NAS with Gradient Matching
Shoukang Hu,Ruochen Wang,Lanqing HONG,Zhenguo Li,Cho-Jui Hsieh,Jiashi Feng
Efficient performance estimation of architectures drawn from large search spaces is essential to Neural Architecture Search. One-Shot methods tackle this challenge by training one supernet to approximate the performance of every architecture in the search space via weight-sharing, thereby drastically reducing the search cost. However, due to coupled optimization between child architectures caused by weight-sharing, One-Shot supernet's performance estimation could be inaccurate, leading to degraded search outcomes. To address this issue, Few-Shot NAS reduces the level of weight-sharing by splitting the One-Shot supernet into multiple separated sub-supernets via edge-wise (layer-wise) exhaustive partitioning. Since each partition of the supernet is not equally important, it necessitates the design of a more effective splitting criterion. In this work, we propose a gradient matching score (GM) that leverages gradient information at the shared weight for making informed splitting decisions. Intuitively, gradients from different child models can be used to identify whether they agree on how to update the shared modules, and subsequently to decide if they should share weight. Compared with exhaustive partitioning, the proposed criterion significantly reduces the branching factor per edge. This allows us to split more edges (layers) for a given budget, resulting in substantially improved performance as NAS search spaces usually include dozens of edges (layers). Extensive empirical evaluations of the proposed method on a wide range of search spaces (NASBench-201, DARTS, MobileNet Space), datasets (cifar10, cifar100, ImageNet) and search algorithms (DARTS, SNAS

, RSPS, ProxylessNAS, OFA) demonstrate that it significantly outperforms its Few-Shot counterparts while surpassing previous comparable methods in terms of the accuracy of derived architectures.
Our code is available at https://github.com/skhu101/GM-NAS.

****

Learnability and Expressiveness in Self-Supervised Learning
Yuchen Lu,Zhen Liu,Alessandro Sordoni,Aristide Baratin,Romain Laroche,Aaron Courville
In this work, we argue that representations induced by self-supervised learning (SSL) methods should both be expressive and learnable. To measure expressiveness, we propose to use the Intrinsic Dimension (ID) of the dataset in representation space. Inspired by the human study of Laina et al. (2020), we introduce Cluster Learnability (CL), defined in terms of the learning speed of a KNN classifier trained to predict K-means cluster labels for held-out representations. By collecting 30 state-of-art checkpoints, both supervised and self-supervised, using different architectures, we show that ID and CL can be combined to predict downstream classification performance better than the existing techniques based on contrastive losses or pretext tasks, while having no requirements on data augmentation, model architecture or human labels. To further demonstrate the utility of our framework, we propose modifying DeepCluster (Caron et al., 2018) to improve the learnability of the representations. Using our modification, we are able to outperform DeepCluster on both STL10 and ImageNet benchmarks. The performance of the intermediate checkpoints can also be well predicted under our framework, suggesting the possibility of developing new SSL algorithms without labels.

****

The Unreasonable Effectiveness of Random Pruning: Return of the Most Naive Baseline for Sparse Training
Shiwei Liu,Tianlong Chen,Xiaohan Chen,Li Shen,Decebal Constantin Mocanu,Zhangyang Wang,Mykola Pechenizkiy
Random pruning is arguably the most naive way to attain sparsity in neural networks, but has been deemed uncompetitive by either post-training pruning or sparse training. In this paper, we focus on sparse training and highlight a perhaps counter-intuitive finding, that random pruning at initialization can be quite powerful for the sparse training of modern neural networks. Without any delicate pruning criteria or carefully pursued sparsity structures, we empirically demonstrate that sparsely training a randomly pruned network from scratch can match the performance of its dense equivalent. There are two key factors that contribute to this revival: (i) $the network sizes matter$: as the original dense networks grow wider and deeper, the performance of training a randomly pruned sparse network will quickly grow to matching that of its dense equivalent, even at high sparsity ratios; (ii) $appropriate layer-wise sparsity ratios$ can be pre-chosen for sparse training, which shows to be another important performance booster. Simple as it looks, a randomly pruned subnetwork of Wide ResNet-50 can be sparsely trained to outperforming a dense Wide ResNet-50, on ImageNet. We also observed such randomly pruned networks outperform dense counterparts in other favorable aspects, such as out-of-distribution detection, uncertainty estimation, and adversarial robustness. Overall, our results strongly suggest there is larger-than-expected room for sparse training at scale, and the benefits of sparsity might be more universal beyond carefully designed pruning. Our source code can be found at https://github.com/VITA-Group/Random_Pruning.

****

Anarchic Federated Learning
Haibo Yang,Xin Zhang,Prashant Khanduri,Jia Liu
Present-day federated learning (FL) systems deployed over edge networks consists of a large number of workers with high degrees of heterogeneity in data and/or computing capabilities, which call for flexible worker participation in terms of timing, effort, data heterogeneity, etc. To achieve these goals, in this work, we propose a new FL paradigm called ``Anarchic Federated Learning'' (AFL). In stark contrast to conventional FL models, each worker in AFL has complete freedom

to choose i) when to participate in FL, and ii) the number of local steps to perform in each round based on its current situation (e.g., battery level, communication channels, privacy concerns). However, AFL also introduces significant challenges in algorithmic design because the server needs to handle the chaotic worker behaviors. Toward this end, we propose two Anarchic Federated Averaging (AFA) algorithms with two-sided learning rates for both cross-device and cross-silo settings, which are named AFA-CD and AFA-CS, respectively. Somewhat surprisingly, even with general worker information arrival processes, we show that both AFL algorithms achieve the same convergence rate order as the state-of-the-art algorithms for conventional FL. Moreover, they retain the highly desirable {\em linear speedup effect} in the new AFL paradigm. We validate the proposed algorithms with extensive experiments on real-world datasets.

****

Weakly Supervised Graph Clustering

Tian Bian,Tingyang Xu,Yu Rong,Wenbing Huang,Xi Xiao,Peilin Zhao,Junzhou Huang,Hong Cheng

Graph Clustering, which clusters the nodes of a graph given its collection of node features and edge connections in an unsupervised manner, has long been researched in graph learning and is essential in certain applications. While this task is common, more complex cases arise in practice—can we cluster nodes better with some graph-level side information or in a weakly supervised manner as, for example, identifying potential fraud users in a social network given additional labels of fraud communities. This triggers an interesting problem which we define as Weakly Supervised Graph Clustering (WSGC). In this paper, we firstly discuss the various possible settings of WSGC, formally. Upon such discussion, we investigate a particular task of weakly supervised graph clustering by making use of the graph labels and node features, with the assistance of a hierarchical graph that further characterizes the connections between different graphs. To address this task, we propose Gaussian Mixture Graph Convolutional Network (GMGCN), a simple yet effective framework for learning node representations under the supervision of graph labels guided by a proposed consensus loss and then inferring the category of each node via a Gaussian Mixture Layer (GML). Extensive experiments are conducted to test the rationality of the formulation of weakly supervised graph clustering. The experimental results show that, with the assistance of graph labels, the weakly supervised graph clustering method has a great improvement over the traditional graph clustering method.

****

On the regularization landscape for the linear recommendation models

Dong Li,Zhenming Liu,Ruoming Jin,Zhi Liu,Jing Gao,Bin Ren

Recently, a wide range of recommendation algorithms inspired by deep learning techniques have emerged as the performance leaders several standard recommendation benchmarks. While these algorithms were built on different DL techniques (e.g., dropouts, autoencoder), they have similar performance and even similar cost functions. This paper studies whether the models' comparable performance are sheer coincidence, or they can be unified into a single framework. We find that all linear performance leaders effectively add only a nuclear-norm based regularizer, or a Frobenius-norm based regularizer. The former ones possess a (surprisnig) rigid structure that limits the models' predictive power but their solutions are low rank and have closed form. The latter ones are more expressive and more efficient for recommendation but their solutions are either full-rank or require executing hard-to-tune numeric procedures such as ADMM. Along this line of finding, we further propose two low-rank, closed-form solutions, derived from carefully generalizing Frobenius-norm based regularizers. The new solutions get the best of both nuclear-norm and Frobenius-norm world.

****

Self-Supervision Enhanced Feature Selection with Correlated Gates

Changhee Lee,Fergus Imrie,Mihaela van der Schaar

Discovering relevant input features for predicting a target variable is a key scientific question. However, in many domains, such as medicine and biology, feature selection is confounded by a scarcity of labeled samples coupled with signifi cant correlations among features. In this paper, we propose a novel deep learning approach to feature selection that addresses both challenges simultaneously. First, we pre-train the network using unlabeled samples within a self-supervised learning framework by solving pretext tasks that require the network to learn informative representations from partial feature sets. Then, we fine-tune the pre-trained network to discover relevant features using labeled samples. During both training phases, we explicitly account for the correlation structure of the input features by generating correlated gate vectors from a multivariate Bernoulli distribution. Experiments on multiple real-world datasets including clinical and omics demonstrate that our model discovers relevant features that provide superior prediction performance compared to the state-of-the-art benchmarks in practical scenarios where there is often limited labeled data and high correlations among features.

****

Score-Based Generative Modeling with Critically-Damped Langevin Diffusion
Tim Dockhorn,Arash Vahdat,Karsten Kreis

Score-based generative models (SGMs) have demonstrated remarkable synthesis quality. SGMs rely on a diffusion process that gradually perturbs the data towards a tractable distribution, while the generative model learns to denoise. The complexity of this denoising task is, apart from the data distribution itself, uniquely determined by the diffusion process. We argue that current SGMs employ overly simplistic diffusions, leading to unnecessarily complex denoising processes, which limit generative modeling performance. Based on connections to statistical mechanics, we propose a novel critically-damped Langevin diffusion (CLD) and show that CLD-based SGMs achieve superior performance. CLD can be interpreted as running a joint diffusion in an extended space, where the auxiliary variables can be considered "velocities" that are coupled to the data variables as in Hamiltonian dynamics. We derive a novel score matching objective for CLD and show that the model only needs to learn the score function of the conditional distribution of the velocity given data, an easier task than learning scores of the data directly. We also derive a new sampling scheme for efficient synthesis from CLD-based diffusion models. We find that CLD outperforms previous SGMs in synthesis quality for similar network architectures and sampling compute budgets. We show that our novel sampler for CLD significantly outperforms solvers such as Euler–Maruyama. Our framework provides new insights into score-based denoising diffusion models and can be readily used for high-resolution image synthesis. Project page and code: https://nv-tlabs.github.io/CLD-SGM.

****

Simpler Calibration for Survival Analysis
Hiroki Yanagisawa,Toshiya Iwamori,Akira Koseki,Michiharu Kudo,Mohamed Ghalwash,Prithwish Chakraborty

Survival analysis, also known as time-to-event analysis, is the problem to predict the distribution of the time of the occurrence of an event. This problem has applications in various fields such as healthcare, security, and finance. While there have been many neural network models proposed for survival analysis, none of them are calibrated. This means that the average of the predicted distribution is different from the actual distribution in the dataset. Therefore, X-CAL has recently been proposed for the calibration, which is supposed to be used as a regularization term in the loss function of a neural network. X-CAL is formulated on the basis of the widely used definition of calibration for distribution regression. In this work, we propose new calibration definitions for distribution regression and survival analysis, and demonstrate a simpler alternative to X-CAL based on the new calibration definition for survival analysis.

****

switch-GLAT: Multilingual Parallel Machine Translation Via Code-Switch Decoder
Zhenqiao Song,Hao Zhou,Lihua Qian,Jingjing Xu,Shanbo Cheng,Mingxuan Wang,Lei Li

Multilingual machine translation aims to develop a single model for multiple language directions. However, existing multilingual models based on Transformer are limited in terms of both translation performance and inference speed. In this p aper, we propose switch-GLAT, a non-autoregressive multilingual machine translation model with a code-switch decoder. It can generate contextual code-switched translations for a given source sentence, and perform code-switch back-translation, greatly boosting multilingual translation performance. In addition, its inference is highly efficient thanks to its parallel decoder. Experiments show that our proposed switch-GLAT outperform the multilingual Transformer with as much as 0.74 BLEU improvement and 6.2x faster decoding speed in inference.

****

TIME-LAPSE: Learning to say “I don't know” through spatio-temporal uncertainty scoring

Nandita Bhaskhar,Daniel Rubin,Christopher Lee-Messer

Safe deployment of trained ML models requires determining when input samples go out-of-distribution (OOD) and refraining from making uncertain predictions on them. Existing approaches inspect test samples in isolation to estimate their corresponding predictive uncertainty. However, in the real-world, deployed models typically see test inputs consecutively and predict labels continuously over time during inference. In this work, we propose TIME-LAPSE, a spatio-temporal framework for uncertainty scoring that examines the sequence of predictions prior to the current sample to determine its predictive uncertainty. Our key insight is that in-distribution samples will be more “similar” to each other compared to OOD samples, not just over the encoding latent-space but also across time. Specifically, (a) our spatial uncertainty score estimates how different OOD latent-space representations are from those of an in-distribution set using metrics such as Mahalanobis distance and cosine similarity and (b) our temporal uncertainty score determines deviations in correlations over time using representations of past inputs in a non-parametric, sliding-window based algorithm. We evaluate TIME-LAPSE on both audio and vision tasks using public datasets and further benchmark our approach on a challenging, real-world, electroencephalograms (EEG) dataset for seizure detection. We achieve state-of-the-art results for OOD detection in the audio and EEG domain and observe considerable gains in semantically corrected vision benchmarks. We show that TIME-LAPSE is more driven by semantic content compared to other methods, i.e., it is more robust to dataset statistics. We also propose a sequential OOD detection evaluation framework to emulate real-life drift settings and show that TIME-LAPSE outperforms spatial methods significantly.

****

Representation Consolidation from Multiple Expert Teachers

Zhizhong Li,Avinash Ravichandran,Charless Fowlkes,Marzia Polito,Rahul Bhotika,Stefano Soatto

A library of diverse expert models transfers better to a novel task than a single generalist model. However, growing such a library indefinitely is impractical. Hence, we explore the problem of learning a consolidated image feature representation from a collection of related task-specific teachers that transfer well on novel recognition tasks. This differs from traditional knowledge distillation in which a student model is trained to emulate the input/output functionality of a teacher. Indeed, we observe experimentally that standard distillation of task-specific teachers, or using these teacher representations directly, **reduces** downstream transferability compared to a task-agnostic generalist model. We show that a simple multi-head, multi-task distillation method using an unlabeled proxy dataset and adding a generalist teacher is sufficient to consolidate representations from task-specific teacher(s). We improve downstream performance, outperforming the teacher (or best of all teachers) as well as the strong baseline of ImageNet pre-trained features. Our method almost reaches the performance of a multi-task joint training oracle, reaping the benefit of the teachers without replaying their training data.

****

Controlling Directions Orthogonal to a Classifier

Yilun Xu,Hao He,Tianxiao Shen,Tommi S. Jaakkola

We propose to identify directions invariant to a given classifier so that these directions can be controlled in tasks such as style transfer. While orthogonal d ecomposition is directly identifiable when the given classifier is linear, we formally define a notion of orthogonality in the non-linear case. We also provide a surprisingly simple method for constructing the orthogonal classifier (a classifier utilizing directions other than those of the given classifier). Empirically, we present three use cases where controlling orthogonal variation is important: style transfer, domain adaptation, and fairness. The orthogonal classifier enables desired style transfer when domains vary in multiple aspects, improves domain adaptation with label shifts and mitigates the unfairness as a predictor. The code is available at https://github.com/Newbeeer/orthogonal_classifier

****

Combining Diverse Feature Priors

Saachi Jain,Dimitris Tsipras,Aleksander Madry

To improve model generalization, model designers often restrict the features that their models use, either implicitly or explicitly. In this work, we explore the design space of leveraging such feature priors by viewing them as distinct perspectives on the data. Specifically, we find that models trained with diverse sets of explicit feature priors have less overlapping failure modes, and can thus be combined more effectively. Moreover, we demonstrate that jointly training such models on additional (unlabeled) data allows them to correct each other's mistakes, which, in turn, leads to better generalization and resilience to spurious correlations.

****

DictFormer: Tiny Transformer with Shared Dictionary

Qian Lou,Ting Hua,Yen-Chang Hsu,Yilin Shen,Hongxia Jin

We introduce DictFormer with the efficient shared dictionary to provide a compact, fast, and accurate transformer model. DictFormer significantly reduces the redundancy in the transformer's parameters by replacing the prior transformer's parameters with a compact, shared dictionary, few unshared coefficients, and indices. Also, DictFormer enables faster computations since expensive weights multiplications are converted into cheap shared look-ups on dictionary and few linear projections. Training dictionary and coefficients are not trivial since indices used for looking up dictionary are not differentiable. We adopt a sparse-constraint training with $l_1\,\,norm$ relaxation to learn coefficients and indices in DictFormer. DictFormer is flexible to support different model sizes by dynamically changing dictionary size. Compared to existing lightweight Transformers, DictFormer consistently reduces model size over Transformer on multiple tasks, e.g., machine translation, abstractive summarization, and language modeling. Extensive experiments show that DictFormer reduces $3.6\times$ to $8.9\times$ model size with similar accuracy over multiple tasks, compared to Transformer.

****

AQUILA: Communication Efficient Federated Learning with Adaptive Quantization of Lazily-Aggregated Gradients

Zihao Zhao,Yuzhu Mao,Muhammad Zeeshan,Yang Liu,Tian Lan,Wenbo Ding

The development and deployment of federated learning (FL) have been bottlenecked by the heavy communication overheads of high-dimensional models between the distributed client nodes and the central server. To achieve better error-communication tradeoffs, recent efforts have been made to either adaptively reduce the communication frequency by skipping unimportant updates, a.k.a. lazily-aggregated quantization (LAQ), or adjust the quantization bits for each communication. In this paper, we propose a unifying communication efficient framework for FL based on adaptive quantization of lazily-aggregated gradients (AQUILA), which adaptively adjusts two mutually-dependent factors, the communication frequency and the quantization level, in a synergistic way. Specifically, we start from a careful investigation on the classical LAQ scheme and formulate AQUILA as an optimization problem where the optimal quantization level per communication is selected by minimizing the gradient loss caused by updates skipping. Meanwhile, we adjust the LAQ strategy to better fit the novel quantization criterion and thus keep the communication frequency at an appropriate level. The effectiveness and convergence of the proposed AQUILA framework are theoretically verified. The experimental results demonstrate that AQUILA can reduce around 50% of overall transmitted bits compared to existing methods while achieving the same level of model accuracy in a number of non-homogeneous FL scenarios, including Non-IID data distribution and heterogeneous model architecture. The proposed AQUILA is highly adaptive and compatible to existing FL settings.

****

Inducing Reusable Skills From Demonstrations with Option-Controller Network
Siyuan Zhou,Yikang Shen,Yuchen Lu,Aaron Courville,Joshua B. Tenenbaum,Chuang Gan

Humans can decompose previous experiences into skills and reuse them to enable fast learning in the future. Inspired by this process, we propose a new model called Option-Controller Network (OCN), which is a bi-level recurrent policy network composed of a high-level controller and a pool of low-level options. The options are disconnected from any task-specific information to model task-agnostic skills.

The controller use options to solve a given task, and it calls one option at a time and waits until the option return. With the isolation of information and the synchronous calling mechanism, we can impose a division of works between the controller and options in an end-to-end training regime. In experiments, we first perform behavior cloning from unstructured demonstrations coming from different tasks. We then freeze the learned options and learn a new controller with an RL algorithm to solve a new task. Extensive results on discrete and continuous environments show that OCN can jointly learn to decompose unstructured demonstrations into skills and model each skill with separate options. The learned options provide a good temporal abstraction, allowing OCN to quickly transfer to tasks with a novel combination of learned skills even with sparse reward, while previous methods either suffer from the delayed reward problem due to the lack of temporal abstraction or a complicated option controlling mechanism that increases the complexity of exploration.

****

Plug-In Inversion: Model-Agnostic Inversion for Vision with Data Augmentations
Amin Ghiasi,Hamid Kazemi,Steven Reich,Chen Zhu,Micah Goldblum,Tom Goldstein

Existing techniques for model inversion typically rely on hard-to-tune regularizers, such as total variation or feature regularization, which must be individually calibrated for each network in order to produce adequate images. In this work, we introduce Plug-In Inversion, which relies on a simple set of augmentations and does not require excessive hyper-parameter tuning. Under our proposed augmentation-based scheme, the same set of augmentation hyper-parameters can be used for inverting a wide range of image classification models, regardless of input dimensions or the architecture. We illustrate the practicality of our approach by inverting Vision Transformers (ViTs) and Multi-Layer Perceptrons (MLPs) trained on the ImageNet dataset, tasks which to the best of our knowledge have not been successfully accomplished by any previous works.

****

Few-Shot Classification with Task-Adaptive Semantic Feature Learning
Meihong Pan,Chunqiu Xia,Hongyi Xin,Yang Yang,Xiaoyong Pan,Hong-Bin Shen

Few-shot classification aims to learn a classifier that categorizes objects of unseen classes with limited samples. One general approach is to mine as much information as possible from limited samples. This can be achieved by incorporating data aspects from multiple modals. However, existing multi-modality methods only use additional modality in support samples while adhering to a single modal in query samples. Such approach could lead to information imbalance between support and query samples, which confounds model generalization from support to query samples. Towards this problem, we propose a task-adaptive semantic feature learning mechanism to incorporates semantic features for both support and query samples. The semantic feature learner is trained episodic-wisely by regressing from the feature vectors of support samples. Then the query samples can obtain the semantic features with this module. Such method maintains a consistent training scheme between support and query samples and enables direct model transfer from support to query datasets, which significantly improves model generalization. We develop two modality combination implementations: feature concatenation and feature

fusion, based on the semantic feature learner. Extensive experiments conducted on four benchmarks demonstrate that our method outperforms state-of-the-arts, proving the effectiveness of our method.
****

Training Transition Policies via Distribution Matching for Complex Tasks
JU-SEUNG BYUN,Andrew Perrault

Humans decompose novel complex tasks into simpler ones to exploit previously learned skills. Analogously, hierarchical reinforcement learning seeks to leverage lower-level policies for simple tasks to solve complex ones. However, because each lower-level policy induces a different distribution of states, transitioning from one lower-level policy to another may fail due to an unexpected starting state. We introduce transition policies that smoothly connect lower-level policies by producing a distribution of states and actions that matches what is expected by the next policy. Training transition policies is challenging because the natural reward signal---whether the next policy can execute its subtask successfully---is sparse. By training transition policies via adversarial inverse reinforcement learning to match the distribution of expected states and actions, we avoid relying on task-based reward. To further improve performance, we use deep Q-learning with a binary action space to determine when to switch from a transition policy to the next pre-trained policy, using the success or failure of the next subtask as the reward. Although the reward is still sparse, the problem is less severe due to the simple binary action space. We demonstrate our method on continuous bipedal locomotion and arm manipulation tasks that require diverse skills. We show that it smoothly connects the lower-level policies, achieving higher success rates than previous methods that search for successful trajectories based on a reward function, but do not match the state distribution.

****

NODEAttack: Adversarial Attack on the Energy Consumption of Neural ODEs
Mirazul Haque,Simin Chen,Wasif Arman Haque,Cong Liu,Wei Yang

Recently, Neural ODE (Ordinary Differential Equation) models have been proposed, which use ordinary differential equation solving to predict the output of neural network. Due to the low memory usage, Neural ODE models can be considered as an alternative that can be deployed in resource-constrained devices (e.g., IoT devices, mobile devices). However, to deploy a Deep Learning model in resource-constrained devices, low inference energy cost is also required along with low memory cost. Unlike the memory cost, the energy consumption of the Neural ODEs during inference can be adaptive because of the adaptive nature of the ODE solvers. Attackers can leverage the adaptive behaviour of Neural ODEs to attack the energy consumption of Neural ODEs. However, energy-based attack scenarios have not been explored against Neural ODEs. To show the vulnerability of Neural ODEs against adversarial energy-based attack, we propose NODEAttack.

The objective of NODEAttack is to generate adversarial inputs that require more ODE solvers computations, therefore increasing neural ODEs inference-time energy consumption.

Our extensive evaluation on two datasets and two popular ODE solvers show that the samples generated through NODEAttack can increase up to 168% energy consumption than average energy consumption of benign test data during inference time. Our evaluation also shows the attack transferability is feasible across solvers and architectures.

Also, we perform a case study showing the impact of the generated adversarial examples, which shows that NODEAttack generated adversarial examples can decrease 50% efficiency of an object-recognition-based mobile application.
****

On Transportation of Mini-batches: A Hierarchical Approach
Khai Nguyen,Dang Nguyen,Nguyen Dinh Quoc,Tung Pham,Hung Bui,Dinh Phung,Trung Le,Nhat Ho

Mini-batch optimal transport (m-OT) has been successfully used in practical applications that involve probability measures with a very high number of supports. The m-OT solves several smaller optimal transport problems and then returns the average of their costs and transportation plans. Despite its scalability advantage, the m-OT does not consider the relationship between mini-batches which leads to undesirable estimation. Moreover, the m-OT does not approximate a proper metric between probability measures since the identity property is not satisfied. To address these problems, we propose a novel mini-batching scheme for optimal transport, named Batch of Mini-batches Optimal Transport (BoMb-OT), that finds the optimal coupling between mini-batches and it can be seen as an approximation to a well-defined distance on the space of probability measures. Furthermore, we show that the m-OT is a limit of the entropic regularized version of the BoMb-OT when the regularized parameter goes to infinity. Finally, we present the new algorithms of the BoMb-OT in various applications, such as deep generative models and deep domain adaptation. From extensive experiments, we observe that the BoMb-OT achieves a favorable performance in deep learning models such as deep generative models and deep domain adaptation. In other applications such as approximate Bayesian computation, color transfer, and gradient flow, the BoMb-OT also yields either a lower quantitative result or a better qualitative result than the m-OT.

****

A Simple and Debiased Sampling Method for Personalized Ranking

Lu Yu,Shichao Pei,Chuxu Zhang,Xiangliang Zhang

Pairwise ranking models have been widely used to address various problems, such as recommendation. The basic idea is to learn the rank of users' preferred items through separating items into positive samples if user-item interactions exist, and negative samples otherwise. Due to the limited number of observed interactions, pairwise ranking models face serious class-imbalance issue. Our theoretical analysis shows that current sampling-based methods cause the vertex-level imbalance problem, which makes the norm of learned item embeddings towards infinite after a certain training iterations, and consequently results in vanishing gradient and affects the model performance. To this end, we propose VINS, an efficient \emph{\underline{Vi}tal \underline{N}egative \underline{S}ampler}, to alleviate the class-imbalance issue for pairwise ranking models optimized by gradient methods. The core of VINS is a bias sampler with reject probability that will tend to accept a negative candidate with a larger popularity than the given positive item. Evaluation results on several real datasets demonstrate that the proposed sampling method speeds up the training procedure 30\% to 50\% for ranking models ranging from shallow to deep, while maintaining and even improving the quality of ranking results in top-N item recommendation.

****

Repairing Systematic Outliers by Learning Clean Subspaces in VAEs

Simao Eduardo,Kai Xu,Alfredo Nazabal,Charles Sutton

Data cleaning often comprises outlier detection and data repair.
Systematic errors result from nearly
deterministic transformations that occur repeatedly in the data,
e.g. specific image pixels being set to default values or watermarks.
Consequently, models with enough capacity easily overfit to these
errors, making detection and repair difficult.
Seeing as a systematic outlier is a combination of patterns of a clean instance
and systematic error patterns, our main insight is that inliers can be
modelled by a smaller representation (subspace) in a model than outliers.
By exploiting this,
we propose \emph{Clean Subspace Variational Autoencoder (CLSVAE)},
a novel semi-supervised model for detection and automated repair of
systematic errors.
The main idea is to partition the latent space and model inlier and
outlier patterns separately.
CLSVAE is effective with much less labelled data compared to previous related
models, often with less than 2\% of the data.
We provide experiments using three image datasets in scenarios with
different levels of corruption and labelled set sizes, comparing to relevant baselines.

CLSVAE provides superior repairs without human intervention, e.g. with just 0.25\% of labelled data we see a relative error decrease of 58\% compared to the closest baseline.

****

Stability analysis of SGD through the normalized loss function

Alexandre Lemire Paquin,Brahim Chaib-draa,Philippe Giguère

We prove new generalization bounds for stochastic gradient descent for both the convex and non-convex cases. Our analysis is based on the stability framework. We analyze stability with respect to the normalized version of the loss function used for training. This leads to investigating a form of angle-wise stability instead of euclidean stability in weights. For neural networks, the measure of distance we consider is invariant to rescaling the weights of each layer. Furthermore, we exploit the notion of on-average stability in order to obtain a data-dependent quantity in the bound. This data-dependent quantity is seen to be more favorable when training with larger learning rates in our numerical experiments. This might help to shed some light on why larger learning rates can lead to better generalization in some practical scenarios.

****

R5: Rule Discovery with Reinforced and Recurrent Relational Reasoning

Shengyao Lu,Bang Liu,Keith G Mills,SHANGLING JUI,Di Niu

Systematicity, i.e., the ability to recombine known parts and rules to form new sequences while reasoning over relational data, is critical to machine intelligence. A model with strong systematicity is able to train on small-scale tasks and generalize to large-scale tasks. In this paper, we propose R5, a relational reasoning framework based on reinforcement learning that reasons over relational graph data and explicitly mines underlying compositional logical rules from observations. R5 has strong systematicity and being robust to noisy data. It consists of a policy value network equipped with Monte Carlo Tree Search to perform recurrent relational prediction and a backtrack rewriting mechanism for rule mining. By alternately applying the two components, R5 progressively learns a set of explicit rules from data and performs explainable and generalizable relation prediction. We conduct extensive evaluations on multiple datasets. Experimental results show that R5 outperforms various embedding-based and rule induction baselines on relation prediction tasks while achieving a high recall rate in discovering ground truth rules.

****

Gradient Broadcast Adaptation: Defending against the backdoor attack in pre-trained models

Tianyu Chen,Haoyi Zhou,He Mingrui,Jianxin Li

Pre-trained language models (e.g, BERT, GPT-3) have revolutionized the NLP research and fine-tuning becomes the indispensable step of downstream adaptation. However, the covert attack is the emerging threat to the pre-train-then-fine tuning learning paradigm. The backdoor attack is a typical challenge, which the victim model fails on the trigger-activated samples while behaves normally on others. These backdoors could survive the cascading fine-tuning stage, which continually posing the application of pre-trained models. In this paper, we proposed a Gradient Broadcast Adaptation (GBA) method, prevent the model from controlled producing outputs in a trigger-anchor-free manner. We design the prompt-based tuning, flexibly accessing the rare tokens while providing a fair measure of distance in word embedding space. The gradient broadcast alleviates lazy updating of potential triggers and purges the underlying abnormal weights. The GBA defense method is evaluated over five text-classification tasks against three state-of-the-art backdoor attacks. We find our method can cover nearly 100% embedded backdoor with negligible performance loss on clean data.

****

GDA-AM: ON THE EFFECTIVENESS OF SOLVING MIN-IMAX OPTIMIZATION VIA ANDERSON MIXING

Huan He,Shifan Zhao,Yuanzhe Xi,Joyce Ho,Yousef Saad

Many modern machine learning algorithms such as generative adversarial networks (GANs) and adversarial training can be formulated as minimax optimization.Gradie nt descent ascent (GDA) is the most commonly used algorithm due to its simplicity. However, GDA can converge to non-optimal minimax points. We propose a new minimax optimization framework,GDA-AM, that views the GDA dynamics as a fixed-point iteration and solves it using Anderson Mixing to converge to the local minimax. It addresses the diverging issue of simultaneous GDA and accelerates the convergence of alternating GDA. We show theoretically that the algorithm can achieve global convergence for bilinear problems under mildconditions. We also empirically show that GDA-AM solves a variety of minimax problems and improves GAN training on several datasets

****

On feature learning in neural networks with global convergence guarantees

Zhengdao Chen,Eric Vanden-Eijnden,Joan Bruna

We study the gradient flow optimization of over-parameterized neural networks (NNs) in a setup that allows feature learning while admitting non-asymptotic global convergence guarantees. First, we prove that for wide shallow NNs under the mean-field (MF) scaling and with a general class of activation functions, when the input dimension is at least the size of the training set, the training loss converges to zero at a linear rate under gradient flow. Building upon this analysis, we study a model of wide multi-layer NNs with random and untrained weights in earlier layers, and also prove a linear-rate convergence of the training loss to zero, regardless of the input dimension. We also show empirically that, unlike in the Neural Tangent Kernel (NTK) regime, our multi-layer model exhibits feature learning and can achieve better generalization performance than its NTK counterpart.

****

The Three Stages of Learning Dynamics in High-dimensional Kernel Methods

Nikhil Ghosh,Song Mei,Bin Yu

To understand how deep learning works, it is crucial to understand the training dynamics of neural networks. Several interesting hypotheses about these dynamics have been made based on empirically observed phenomena, but there exists a limited theoretical understanding of when and why such phenomena occur.

In this paper, we consider the training dynamics of gradient flow on kernel least-squares objectives, which is a limiting dynamics of SGD trained neural networks. Using precise high-dimensional asymptotics, we characterize the dynamics of the fitted model in two “worlds”: in the Oracle World the model is trained on the population distribution and in the Empirical World the model is trained on an i.i.d finite dataset. We show that under mild conditions on the kernel and L^2 target regression function the training dynamics have three stages that are based on the behaviors of the models in the two worlds. Our theoretical results also mathematically formalize some interesting deep learning phenomena. Specifically, in our setting we show that SGD progressively learns more complex functions and that there is a "deep bootstrap" phenomenon: during the second stage, the test error of both worlds remain close despite the empirical training error being much smaller. Finally, we give a concrete example comparing the dynamics of two different kernels which shows that faster training is not necessary for better generalization.

****

Eigenspace Restructuring: a Principle of Space and Frequency in Neural Networks

Lechao Xiao

Understanding the fundamental principles behind the massive success of neural networks is one of the most important open questions in deep learning. However, due to the highly complex nature of the problem, progress has been relatively slow.In this note, through the lens of infinite-width networks, a.k.a. neural kernels, we present one such principle resulting from hierarchical locality. It is well-known that the eigenstructure of infinite-width multilayer perceptrons (MLPs) depends solely on the concept frequency, which measures the order of interactions. We show that the topologies from convolutional networks (CNNs) restructure the associated eigenspaces into finer subspaces. In addition to frequency, the new structure also depends on the concept space— the distance among interaction te rms, defined via the length of a minimum spanning tree containing them. The resulting fine-grained eigenstructure dramatically improves the network’s learnability, empowering them to simultaneously model a much richer class of interactions, including long-range-low-frequency interactions, short-range-high-frequency interactions, and various interpolations and extrapolations in-between. Finally, we show that increasing the depth of a CNN can improve the inter/extrapolation resolution and, therefore, the network’s learnability.

****

When Can We Learn General-Sum Markov Games with a Large Number of Players Sample-Efficiently?

Ziang Song,Song Mei,Yu Bai

Multi-agent reinforcement learning has made substantial empirical progresses in solving games with a large number of players. However, theoretically, the best known sample complexity for finding a Nash equilibrium in general-sum games scales exponentially in the number of players due to the size of the joint action space, and there is a matching exponential lower bound. This paper investigates what learning goals admit better sample complexities in the setting of m-player general-sum Markov games with H steps, S states, and A_i actions per player. First, we design algorithms for learning an ϵ-Coarse Correlated Equilibrium (CCE) in $\widetilde{\mathcal{O}}(H^5S\max_{i\le m} A_i / \epsilon^2)$ episodes, and an ϵ-Correlated Equilibrium (CE) in $\widetilde{\mathcal{O}}(H^6S\max_{i\le m} A_i^2 / \epsilon^2)$ episodes. This is the first line of results for learning CCE and CE with sample complexities polynomial in $\max_{i\le m} A_i$. Our algorithm for learning CE integrates an adversarial bandit subroutine which minimizes a weighted swap regret, along with several novel designs in the outer loop. Second, we consider the important special case of Markov Potential Games, and design an algorithm that learns an ϵ-approximate Nash equilibrium within $\widetilde{\mathcal{O}}(S\sum_{i\le m} A_i / \epsilon^3)$ episodes (when only highlighting the dependence on S, A_i, and ϵ), which only depends linearly in $\sum_{i\le m} A_i$ and significantly improves over the existing efficient algorithm in the ϵ dependence. Overall, our results shed light on what equilibria or structural assumptions on the game may enable sample-efficient learning with many players.

****

Theoretical Analysis of Consistency Regularization with Limited Augmented Data

Shuo Yang,Yijun Dong,Rachel Ward,Inderjit S Dhillon,sujay sanghavi,Qi Lei

Data augmentation is popular in the training of large neural networks; currently, however, there is no clear theoretical comparison between different algorithmic choices on how to use augmented data. In this paper, we take a small step in this direction; we present a simple new statistical framework to analyze data augmentation - specifically, one that captures what it means for one input sample to be an augmentation of another, and also the richness of the augmented set. We use this to interpret consistency regularization as a way to reduce function class complexity, and characterize its generalization performance. Specializing this analysis for linear regression shows that consistency regularization has strictly better sample efficiency as compared to empirical risk minimization on the augmented set. In addition, we also provide generalization bounds under consistency regularization for logistic regression and two-layer neural networks. We perform experiments that make a clean and apples-to-apples comparison (i.e. with no extra modeling or data tweaks) between ERM and consistency regularization using CIFAR-100 and WideResNet; these demonstrate the superior efficacy of consistency regularization.

****

Representation Learning for Online and Offline RL in Low-rank MDPs

Masatoshi Uehara,Xuezhou Zhang,Wen Sun

This work studies the question of Representation Learning in RL: how can we learn a compact low-dimensional representation such that on top of the representation we can perform RL procedures such as exploration and exploitation, in a sample efficient manner. We focus on the low-rank Markov Decision Processes (MDPs) where the transition dynamics correspond to a low-rank transition matrix. Unlike pr ior works that assume the representation is known (e.g., linear MDPs), here we need to learn the representation for the low-rank MDP. We study both the online RL and offline RL settings. For the online setting, operating with the same computational oracles used in FLAMBE (Agarwal et.al), the state-of-art algorithm for learning representations in low-rank MDPs, we propose an algorithm REP-UCB Upper Confidence Bound driven Representation learning for RL), which significantly improves the sample complexity from $\widetilde{O}(A^9 d^7 / (\epsilon^{10} (1-\gamma)^{22}))$ for FLAMBE to $\widetilde{O}(A^4 d^4 / (\epsilon^2 (1-\gamma)^{2}))$ with d being the rank of the transition matrix (or dimension of the ground truth representation), A being the number of actions, and γ being the discounted factor. Notably, REP-UCB is simpler than FLAMBE, as it directly balances the interplay between representation learning, exploration, and exploitation, while FLAMBE is an explore-then-commit style approach and has to perform reward-free exploration step-by-step forward in time. For the offline RL setting, we develop an algorithm that leverages pessimism to learn under a partial coverage condition: our algorithm is able to compete against any policy as long as it is covered by the offline distribution.

****

Counting Substructures with Higher-Order Graph Neural Networks: Possibility and Impossibility Results

Behrooz Tahmasebi,Derek Lim,Stefanie Jegelka

While message passing Graph Neural Networks (GNNs) have become increasingly popular architectures for learning with graphs, recent works have revealed important shortcomings in their expressive power. In response, several higher-order GNNs have been proposed that substantially increase the expressive power, albeit at a large computational cost. Motivated by this gap, we explore alternative strategies and lower bounds. In particular, we analyze a new recursive pooling technique of local neighborhoods that allows different tradeoffs of computational cost and expressive power. First, we prove that this model can count subgraphs of size k, and thereby overcomes a known limitation of low-order GNNs. Second, we show how recursive pooling can exploit sparsity to reduce the computational complexity compared to the existing higher-order GNNs. More generally, we provide a (near) matching information-theoretic lower bound for counting subgraphs with graph representations that pool over representations of derived (sub-)graphs. We also discuss lower bounds on time complexity.

****

Neural Networks as Kernel Learners: The Silent Alignment Effect

Alexander Atanasov,Blake Bordelon,Cengiz Pehlevan

Neural networks in the lazy training regime converge to kernel machines. Can neural networks in the rich feature learning regime learn a kernel machine with a data-dependent kernel? We demonstrate that this can indeed happen due to a phenomenon we term silent alignment, which requires that the tangent kernel of a network evolves in eigenstructure while small and before the loss appreciably decreases, and grows only in overall scale afterwards. We show that such an effect takes place in homogenous neural networks with small initialization and whitened data. We provide an analytical treatment of this effect in the linear network case. In general, we find that the kernel develops a low-rank contribution in the early phase of training, and then evolves in overall scale, yielding a function equivalent to a kernel regression solution with the final network's tangent kernel. The early spectral learning of the kernel depends on the depth. We also demonstrate that non-whitened data can weaken the silent alignment effect.

****

Learning Object-Oriented Dynamics for Planning from Text

Guiliang Liu,Ashutosh Adhikari,Amir-massoud Farahmand,Pascal Poupart

The advancement of dynamics models enables model-based planning in complex environments. Existing dynamics models commonly study image-based games with fully observable states. Generalizing these models to Text-Based Games (TBGs), which commonly describe the partially observable states with noisy text observations, is challenging. In this work, we propose an Object-Oriented Text Dynamics (OOTD) model that enables planning algorithms to solve decision-making problems in text d omains. OOTD predicts a memory graph that dynamically remembers the history of object observations and filters object-irrelevant information. To facilitate the robustness of dynamics, our OOTD model identifies the objects influenced by input actions and predicts the belief of object states with independently parameterized transition layers. We develop variational objectives under the object-supervised and self-supervised settings to model the stochasticity of predicted dynamics. Empirical results show OOTD-based planner significantly outperforms model-free baselines in terms of sample efficiency and running scores.

****

Meta-Learning Dynamics Forecasting Using Task Inference

Rui Wang,Robin Walters,Rose Yu

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 which 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 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 both turbulent flow and real-world ocean data forecasting tasks.

****

Message Function Search for Hyper-relational Knowledge Graph

Shimin Di,Lei Chen

Recently, the hyper-relational knowledge graph (HKG) has attracted much attention due to its widespread existence and potential applications. The pioneer works have adapted powerful graph neural networks (GNNs) to embed HKGs by proposing domain-specific message functions. These message functions for HKG embedding are utilized to learn relational representations and capture the correlation between entities and relations of HKGs. However, these works often manually design and fix structures and operators of message functions, which makes them difficult to handle complex and diverse relational patterns in various HKGs (i.e., data patterns). To overcome these shortcomings, we plan to develop a method to dynamically search suitable message functions that can adapt to patterns of the given HKG. Unfortunately, it is not trivial to design an expressive search space and an efficient search algorithm to make the search effective and efficient. In this paper, we first unify a search space of message functions that enables both structures and operators to be searchable. Especially, the classic KG/HKG models and message functions of existing GNNs can be instantiated as special cases in the proposed search space. Then, we design an efficient search algorithm to search the message function and other GNN components for any given HKGs. Through empirical study, we show that the searched message functions are data-dependent, and can achieve leading performance in link/relation prediction tasks on benchmark data sets.

****

Transformers are Meta-Reinforcement Learners

Luckeciano Carvalho Melo

The transformer architecture and variants presented a remarkable success across many machine learning tasks in recent years. This success is intrinsically related to the capability of handling long sequences and the presence of context-dependent weights from the attention mechanism. We argue that these capabilities suit the central role of a Meta-Reinforcement Learning algorithm. Indeed, a meta-RL agent needs to infer the task from a sequence of trajectories. Furthermore, it requires a fast adaptation strategy to adapt its policy for a new task - which can be achieved using the self-attention mechanism. In this work, we present TrMRL (Transformers for Meta-Reinforcement Learning), a meta-RL agent that mimics th

e memory reinstatement mechanism using the transformer architecture. It associates the recent past of working memories to build an episodic memory recursively through the transformer layers. This memory works as a proxy to the current task, and we condition a policy head on it. We conducted experiments in high-dimensional continuous control environments for locomotion and dexterous manipulation. Results show that TrMRL achieves or surpasses state-of-the-art performance, sample efficiency, and out-of-distribution generalization in these environments.
****

An Operator Theoretic View On Pruning Deep Neural Networks
William T Redman,MARIA FONOBEROVA,Ryan Mohr,Yannis Kevrekidis,Igor Mezic
The discovery of sparse subnetworks that are able to perform as well as full models has found broad applied and theoretical interest. While many pruning methods have been developed to this end, the naïve approach of removing parameters based on their magnitude has been found to be as robust as more complex, state-of-the-art algorithms. The lack of theory behind magnitude pruning's success, especially pre-convergence, and its relation to other pruning methods, such as gradient based pruning, are outstanding open questions in the field that are in need of being addressed. We make use of recent advances in dynamical systems theory, namely Koopman operator theory, to define a new class of theoretically motivated pruning algorithms. We show that these algorithms can be equivalent to magnitude and gradient based pruning, unifying these seemingly disparate methods, and find that they can be used to shed light on magnitude pruning's performance during the early part of training.
****

Lossless Compression with Probabilistic Circuits
Anji Liu,Stephan Mandt,Guy Van den Broeck
Despite extensive progress on image generation, common deep generative model architectures are not easily applied to lossless compression. For example, VAEs suffer from a compression cost overhead due to their latent variables. This overhead can only be partially eliminated with elaborate schemes such as bits-back coding, often resulting in poor single-sample compression rates. To overcome such problems, we establish a new class of tractable lossless compression models that permit efficient encoding and decoding: Probabilistic Circuits (PCs). These are a class of neural networks involving $|p|$ computational units that support efficient marginalization over arbitrary subsets of the D feature dimensions, enabling efficient arithmetic coding. We derive efficient encoding and decoding schemes that both have time complexity $\mathcal{O} (\log(D) \cdot |p|)$, where a naive scheme would have linear costs in D and $|p|$, making the approach highly scalable. Empirically, our PC-based (de)compression algorithm runs 5-40 times faster than neural compression algorithms that achieve similar bitrates. By scaling up the traditional PC structure learning pipeline, we achieve state-of-the-art results on image datasets such as MNIST. Furthermore, PCs can be naturally integrated with existing neural compression algorithms to improve the performance of these base models on natural image datasets. Our results highlight the potential impact that non-standard learning architectures may have on neural data compression.
****

GARNET: A Spectral Approach to Robust and Scalable Graph Neural Networks
Chenhui Deng,Xiuyu Li,Zhuo Feng,Zhiru Zhang
Graph neural networks (GNNs) have been increasingly deployed in various applications that involve learning on non-Euclidean data. However, recent studies show that GNNs are vulnerable to graph adversarial attacks. Although there are several defense methods to improve GNN adversarial robustness, they fail to perform well on low homophily graphs. In addition, few of those defense models can scale to large graphs due to their high computational complexity and memory usage. In this paper, we propose GARNET, a scalable spectral method to boost the adversarial robustness of GNN models for both homophilic and heterophilic graphs. GARNET first computes a reduced-rank yet sparse approximation of the adversarial graph by exploiting an efficient spectral graph embedding and sparsification scheme. Next, GARNET trains an adaptive graph filter on the reduced-rank graph for node re

presentation refinement, which is subsequently leveraged to guide label propagation for further enhancing the quality of node embeddings. GARNET has been evaluated on both homophilic and heterophilic datasets, including a large graph with millions of nodes. Our extensive experiment results show that GARNET increases adversarial accuracy over state-of-the-art GNN (defense) models by up to 9.96% and 15.17% on homophilic and heterophilic graphs, respectively.

****

Capacity of Group-invariant Linear Readouts from Equivariant Representations: How Many Objects can be Linearly Classified Under All Possible Views?

Matthew Farrell,Blake Bordelon,Shubhendu Trivedi,Cengiz Pehlevan

Equivariance has emerged as a desirable property of representations of objects subject to identity-preserving transformations that constitute a group, such as translations and rotations. However, the expressivity of a representation constrained by group equivariance is still not fully understood. We address this gap by providing a generalization of Cover's Function Counting Theorem that quantifies the number of linearly separable and group-invariant binary dichotomies that can be assigned to equivariant representations of objects. We find that the fraction of separable dichotomies is determined by the dimension of the space that is fixed by the group action. We show how this relation extends to operations such as convolutions, element-wise nonlinearities, and global and local pooling. While other operations do not change the fraction of separable dichotomies, local pooling decreases the fraction, despite being a highly nonlinear operation. Finally, we test our theory on intermediate representations of randomly initialized and fully trained convolutional neural networks and find perfect agreement.

****

Tuformer: Data-driven Design of Transformers for Improved Generalization or Efficiency

Xiaoyu Liu,Jiahao Su,Furong Huang

Transformers are neural network architectures that achieve remarkable performance in many areas. However, the core component of Transformers, multi-head self-attention (MHSA), is mainly derived from heuristics, and the interactions across its components are not well understood. To address the problem, we first introduce a mathematically rigorous and yet intuitive tensor diagram representation of MHSA. Guided by tensor diagram representations, we propose a novel design, namely Tunable Transformers (Tuformers), by allowing data-driven weights across heads, whereas MHSA adopts pre-defined and fixed weights across heads, as will be explained in our paper. Tuformers naturally reveal a flexible design space that a user, depending on the needs, can choose a structure that has either improved performance (generalization error) or higher model efficiency. Any pre-trained Transformer can be an initialization of the corresponding Tuformer with trainable number of heads for efficient training and fine-tuning. Tuformers universally outperform Transformers on various tasks across multiple domains under a wide range of model sizes.

****

Apollo: An Adaptive Parameter-wised Diagonal Quasi-Newton Method for Nonconvex Stochastic Optimization

Xuezhe Ma

In this paper, we introduce Apollo, a quasi-Newton method for nonconvex stochastic optimization, which dynamically incorporates the curvature of the loss function by approximating the Hessian via a diagonal matrix. Importantly, the update and storage of the diagonal approximation of Hessian is as efficient as adaptive first-order optimization methods with linear complexity for both time and memory. To handle nonconvexity, we replace the Hessian with its rectified absolute value, which is guaranteed to be positive-definite. Experiments on three tasks of vision and language show that Apollo achieves significant improvements over other stochastic optimization methods, including SGD and variants of Adam, in term of both convergence speed and generalization performance.

****

Learning Weakly-supervised Contrastive Representations

Yao-Hung Hubert Tsai,Tianqin Li,Weixin Liu,Peiyuan Liao,Ruslan Salakhutdinov,Lou is-Philippe Morency

We argue that a form of the valuable information provided by the auxiliary information is its implied data clustering information. For instance, considering hashtags as auxiliary information, we can hypothesize that an Instagram image will be semantically more similar with the same hashtags. With this intuition, we present a two-stage weakly-supervised contrastive learning approach. The first stage is to cluster data according to its auxiliary information. The second stage is to learn similar representations within the same cluster and dissimilar representations for data from different clusters. Our empirical experiments suggest the following three contributions. First, compared to conventional self-supervised representations, the auxiliary-information-infused representations bring the performance closer to the supervised representations, which use direct downstream labels as supervision signals. Second, our approach performs the best in most cases, when comparing our approach with other baseline representation learning methods that also leverage auxiliary data information. Third, we show that our approach also works well with unsupervised constructed clusters (e.g., no auxiliary information), resulting in a strong unsupervised representation learning approach.

****

Encoding Weights of Irregular Sparsity for Fixed-to-Fixed Model Compression

Bae Seong Park,Se Jung Kwon,Daehwan Oh,Byeongwook Kim,Dongsoo Lee

Even though fine-grained pruning techniques achieve a high compression ratio, conventional sparsity representations (such as CSR) associated with irregular sparsity degrade parallelism significantly. Practical pruning methods, thus, usually lower pruning rates (by structured pruning) to improve parallelism. In this paper, we study fixed-to-fixed (lossless) encoding architecture/algorithm to support fine-grained pruning methods such that sparse neural networks can be stored in a highly regular structure. We first estimate the maximum compression ratio of encoding-based compression using entropy. Then, as an effort to push the compression ratio to the theoretical maximum (by entropy), we propose a sequential fixed-to-fixed encoding scheme. We demonstrate that our proposed compression scheme achieves almost the maximum compression ratio for the Transformer and ResNet-50 pruned by various fine-grained pruning methods.

****

Semi-supervised Long-tailed Recognition using Alternate Sampling

Bo Liu,Haoxiang Li,Hao Kang,Nuno Vasconcelos,Gang Hua

Main challenges in long-tailed recognition come from the imbalanced data distribution and sample scarcity in its tail classes. While techniques have been proposed to achieve a more balanced training loss and to improve tail classes data variations with synthesized samples, we resort to leverage readily available unlabeled data to boost recognition accuracy. The idea leads to a new recognition setting, namely semi-supervised long-tailed recognition. We argue this setting better resembles the real-world data collection and annotation process and hence can help close the gap to real-world scenarios. To address the semi-supervised long-tailed recognition problem, we present an alternate sampling framework combining the intuitions from successful methods in these two research areas. The classifier and feature embedding are learned separately and updated iteratively. The class-balanced sampling strategy has been implemented to train the classifier in a way not affected by the pseudo labels' quality on the unlabeled data. A consistency loss has been introduced to limit the impact from unlabeled data while leveraging them to update the feature embedding. We demonstrate significant accuracy improvements over other competitive methods on two datasets.

****

A Closer Look at Prototype Classifier for Few-shot Image Classification

Mingcheng Hou,Issei Sato

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 a linear classifier with fine-tuning, which does not use a meta-learning algorithm, performs comparably with the prototypical network.

However, fine-tuning requires additional hyper-parameters when adapting a model to a new environment. In addition, although the purpose of few-shot learning is to enable the model to quickly adapt to a new environment, fine-tuning needs to be applied every time a new class appears, making fast adaptation difficult.

In this paper,

we analyze how a prototype classifier works equally well without fine-tuning and meta-learning.

We experimentally found that directly using the feature vector extracted using standard pre-trained models to construct a prototype classifier in meta-testing does not perform as well as the prototypical network and linear classifiers with fine-tuning and feature vectors of pre-trained models.

Thus, we derive a novel generalization bound for the prototypical network and show that focusing on the variance of the norm of a feature vector can improve performance.

We experimentally investigated several normalization methods for minimizing the variance of the norm and found that the same performance can be obtained by using the L2 normalization and embedding space transformation without fine-tuning or meta-learning.

****

Multi-agent Performative Prediction: From Global Stability and Optimality to Chaos

Georgios Piliouras,Fang-Yi Yu

The recent framework of performative prediction is aimed at capturing settings where predictions influence the target/outcome they try to predict. In this paper, we introduce a natural multi-agent version of this framework, where multiple decision makers try to predict the same outcome. We showcase that such competition can result in interesting phenomena by proving the possibility of phase transitions from stability to instability and eventually chaos. Specifically, we present settings of multi-agent performative prediction where under sufficient conditions, their dynamics lead to global stability and optimality. In the opposite direction, when the agents are not sufficiently cautious in their learning/updates rates, we show that instability and in fact formal chaos is possible. We complement our theoretical predictions with simulations showcasing the predictive power of our results.

****

Dynamic Least-Squares Regression

Binghui Peng,Shunhua Jiang,OMRI WEINSTEIN

In large-scale supervised learning, after a model is trained with an initial dataset, a common challenge is how to exploit new incremental data without re-training the model from scratch. Motivated by this problem, we revisit the canonical problem of dynamic least-squares regression (LSR), where the goal is to learn a linear model over incremental training data. In this setup, data and labels $(\mathbf{A}^{(t)}, \mathbf{b}^{(t)}) \in \mathbb{R}^{t \times d}\times \mathbb{R}^t$ evolve in an online fashion ($t\gg d$), and the goal is to efficiently maintain an (approximate) solution of $\min_{\mathbf{x}^{(t)}} \| \mathbf{A}^{(t)} \mathbf{x}^{(t)} - \mathbf{b}^{(t)} \|_2$ for all $t\in [T]$. Our main result is a dynamic data structure which maintains an arbitrarily small constant approximate solution to dynamic LSR with amortized update time $O(d^{1+o(1)})$, almost matching the running time of the static (sketching-based) solution. By contrast, for exact (or $1/\mathrm{poly}(n)$-accuracy) solutions, we show a separation between the models, namely, that dynamic LSR requires $\Omega(d^{2-o(1)})$ amortized update time under the OMv Conjecture (Henzinger et al., STOC'15). Our data structure is fast, conceptually simple, easy to implement, and our experiments demonstrate their practicality on both synthetic and real-world datasets.

****

An Experimental Design Perspective on Model-Based Reinforcement Learning

Viraj Mehta,Biswajit Paria,Jeff Schneider,Stefano Ermon,Willie Neiswanger

In many practical applications of RL, it is expensive to observe state transitio ns from the environment. For example, in the problem of plasma control for nuclear fusion, computing the next state for a given state-action pair requires querying an expensive transition function which can lead to many hours of computer simulation or dollars of scientific research. Such expensive data collection prohibits application of standard RL algorithms which usually require a large number of observations to learn. In this work, we address the problem of efficiently learning a policy while making a minimal number of state-action queries to the transition function. In particular, we leverage ideas from Bayesian optimal experimental design to guide the selection of state-action queries for efficient learning. We propose an \emph{acquisition function} that quantifies how much information a state-action pair would provide about the optimal solution to a Markov decision process. At each iteration, our algorithm maximizes this acquisition function, to choose the most informative state-action pair to be queried, thus yielding a data-efficient RL approach. We experiment with a variety of simulated continuous control problems and show that our approach learns an optimal policy with up to 5 -- $1,000\times$ less data than model-based RL baselines and 10^3 -- $10^5\times$ less data than model-free RL baselines. We also provide several ablated comparisons which point to substantial improvements arising from the principled method of obtaining data.

****

Conjugation Invariant Learning with Neural Networks

Aaron Yi Rui Low,Subhroshekhar Ghosh,Yong Sheng Soh

Machine learning under the constraint of symmetries, given by group invariances or equivariances, has emerged as a topic of active interest in recent years. Natural settings for such applications include the multi-reference alignment and cryo electron microscopy, multi-object tracking, spherical images, and so on. A fundamental paradigm among such symmetries is the action of a group by symmetries, which often pertains to change of basis or relabelling of objects in pure and applied mathematics. Thus, a naturally significant class of functions consists of those that are intrinsic to the problem, in the sense of being independent of such base change or relabelling; in other words invariant under the conjugation action by a group. In this work, we investigate such functions, known as class functions, leveraging tools from group representation theory. A fundamental ingredient in our approach are given by the so-called irreducible characters of the group, which are canonical tracial class functions related to its irreducible representations. Such functions form an orthogonal basis for the class functions, extending ideas from Fourier analysis to this domain, and accord a very explicit structure. Exploiting a tensorial structure on representations, which translates into a multiplicative algebra structure for irreducible characters, we propose to efficiently approximate class functions using polynomials in a small number of such characters. Thus, our approach provides a global, non-linear coordinate system to describe functions on the group that is intrinsic in nature, in the sense that it is independent of local charts, and can be easily computed in concrete models. We demonstrate that such non-linear approximation using a small dictionary can be effectively implemented using a deep neural network paradigm. This allows us to learn a class function efficiently from a dataset of its outputs.

****

Generate, Annotate, and Learn: Generative Models Advance Self-Training and Knowledge Distillation

Xuanli He,Islam Nassar,Jamie Ryan Kiros,Gholamreza Haffari,Mohammad Norouzi

Semi-Supervised Learning (SSL) has seen success in many application domains, but this success often relies on the availability of task-specific unlabeled data. Knowledge distillation (KD) has enabled compressing deep networks, achieving the best results when distilling knowledge on fresh task-specific unlabeled examples. However, task-specific unlabeled data can be challenging to find, especially for NLP problems. We present a simple framework called "generate, annotate, and learn (GAL)" that uses unconditional language models to synthesize in-domain unlabeled data, helping advance SSL and KD on NLP and tabular tasks. To obtain strong task-specific generative models, we either fine-tune a large language model (LLM) on inputs from specific tasks, or prompt a LLM with a few input examples to generate more unlabeled examples. Then, we use existing classifiers to annotate generated unlabeled examples with pseudo labels, which are used as additional training data or as additional prompts. GAL improves prompt-based few-shot learning on several NLP tasks. It also yields a new state-of-the-art for 6-layer transformers on the GLUE leaderboard. Finally, self-training with GAL offers large gains on four tabular tasks from the UCI repository.

****

BAM: Bayes with Adaptive Memory

Josue Nassar,Jennifer Rogers Brennan,Ben Evans,Kendall Lowrey

Online learning via Bayes' theorem allows new data to be continuously integrated into an agent's current beliefs. However, a naive application of Bayesian methods in non-stationary environments leads to slow adaptation and results in state estimates that may converge confidently to the wrong parameter value. A common solution when learning in changing environments is to discard/downweight past data; however, this simple mechanism of "forgetting" fails to account for the fact that many real-world environments involve revisiting similar states. We propose a new framework, Bayes with Adaptive Memory (BAM), that takes advantage of past experience by allowing the agent to choose which past observations to remember and which to forget. We demonstrate that BAM generalizes many popular Bayesian update rules for non-stationary environments. Through a variety of experiments, we demonstrate the ability of BAM to continuously adapt in an ever-changing world.

****

α-Weighted Federated Adversarial Training

Jianing Zhu,Jiangchao Yao,Tongliang Liu,Kunyang Jia,Jingren Zhou,Bo Han,Hongxia Yang

Federated Adversarial Training (FAT) helps us address the data privacy and governance issues, meanwhile maintains the model robustness to the adversarial attack. However, the inner-maximization optimization of Adversarial Training can exacerbate the data heterogeneity among local clients, which triggers the pain points of Federated Learning. This makes that the straightforward combination of two paradigms shows the performance deterioration as observed in previous works. In this paper, we introduce an α-Weighted Federated Adversarial Training (α-WFAT) method to overcome this problem, which relaxes the inner-maximization of Adversarial Training into a lower bound friendly to Federated Learning. We present the theoretical analysis about this α-weighted mechanism and its effect on the convergence of FAT. Empirically, the extensive experiments are conducted to comprehensively understand the characteristics of α-WFAT, and the results on three benchmark datasets demonstrate α-WFAT significantly outperforms FAT under different adversarial learning methods and federated optimization methods.

****

Learning with convolution and pooling operations in kernel methods

Theodor Misiakiewicz,Song Mei

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 the success of these architectures is that they 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 fully characterize the RKHS of kernels composed of single layers of convolution, pooling, and downsampling operations. We then study the gain in sample efficiency of kernel methods using these kernels over standard inner-product kernels. In particular, we show that 1) the convolution layer breaks the curse of dimensionality by restricting the RKHS to `local' functions; 2) local pooling biases learning towards low-frequency functions, which are stable by small translations; 3) downsampling may modify the high-frequency eigenspaces but leaves the low-frequency part approximately unchanged. Notably, our results quantify how choosing an architecture adapted to the target function leads to a large improvement in the sample complexity.

****

Unsupervised Learning of Full-Waveform Inversion: Connecting CNN and Partial Differential Equation in a Loop

Peng Jin,Xitong Zhang,Yinpeng Chen,Sharon X Huang,Zicheng Liu,Youzuo Lin

This paper investigates unsupervised learning of Full-Waveform Inversion (FWI), which has been widely used in geophysics to estimate subsurface velocity maps from seismic data. This problem is mathematically formulated by a second order partial differential equation (PDE), but is hard to solve. Moreover, acquiring velocity map is extremely expensive, making it impractical to scale up a supervised approach to train the mapping from seismic data to velocity maps with convolutional neural networks (CNN).We address these difficulties by $\textit{integrating PDE and CNN in a loop}$, thus shifting the paradigm to unsupervised learning that only requires seismic data. In particular, we use finite difference to approximate the forward modeling of PDE as a differentiable operator (from velocity map to seismic data) and model its inversion by CNN (from seismic data to velocity map). Hence, we transform the supervised inversion task into an unsupervised seismic data reconstruction task. We also introduce a new large-scale dataset $\textit{OpenFWI}$, to establish a more challenging benchmark for the community. Experiment results show that our model (using seismic data alone) yields comparable accuracy to the supervised counterpart (using both seismic data and velocity map). Furthermore, it outperforms the supervised model when involving more seismic data.

****

Conditional Contrastive Learning with Kernel

Yao-Hung Hubert Tsai,Tianqin Li,Martin Q. Ma,Han Zhao,Kun Zhang,Louis-Philippe Morency,Ruslan Salakhutdinov

Conditional contrastive learning frameworks consider the conditional sampling procedure that constructs positive or negative data pairs conditioned on specific variables. Fair contrastive learning constructs negative pairs, for example, from the same gender (conditioning on sensitive information), which in turn reduces undesirable information from the learned representations; weakly supervised contrastive learning constructs positive pairs with similar annotative attributes (conditioning on auxiliary information), which in turn are incorporated into the representations. Although conditional contrastive learning enables many applications, the conditional sampling procedure can be challenging if we cannot obtain sufficient data pairs for some values of the conditioning variable. This paper presents Conditional Contrastive Learning with Kernel (CCL-K) that converts existing conditional contrastive objectives into alternative forms that mitigate the insufficient data problem. Instead of sampling data according to the value of the conditioning variable, CCL-K uses the Kernel Conditional Embedding Operator that samples data from all available data and assigns weights to each sampled data given the kernel similarity between the values of the conditioning variable. We conduct experiments using weakly supervised, fair, and hard negatives contrastive learning, showing CCL-K outperforms state-of-the-art baselines.

****

Frame Averaging for Invariant and Equivariant Network Design

Omri Puny,Matan Atzmon,Edward J. Smith,Ishan Misra,Aditya Grover,Heli Ben-Hamu,Yaron Lipman

Many machine learning tasks involve learning functions that are known to be invariant or equivariant to certain symmetries of the input data. However, it is often challenging to design neural network architectures that respect these symmetries while being expressive and computationally efficient. For example, Euclidean motion invariant/equivariant graph or point cloud neural networks.

We introduce Frame Averaging (FA), a highly general purpose and systematic framework for adapting known (backbone) architectures to become invariant or equivariant to new symmetry types. Our framework builds on the well known group averaging operator that guarantees invariance or equivariance but is intractable. In con

trast, we observe that for many important classes of symmetries, this operator can be replaced with an averaging operator over a small subset of the group elements, called a frame. We show that averaging over a frame guarantees exact invariance or equivariance while often being much simpler to compute than averaging over the entire group. Furthermore, we prove that FA-based models have maximal expressive power in a broad setting and in general preserve the expressive power of their backbone architectures. Using frame averaging, we propose a new class of universal Graph Neural Networks (GNNs), universal Euclidean motion invariant point cloud networks, and Euclidean motion invariant Message Passing (MP) GNNs. We demonstrate the practical effectiveness of FA on several applications including point cloud normal estimation, beyond 2-WL graph separation, and n-body dynamics prediction, achieving state-of-the-art results in all of these benchmarks.
****

ConFeSS: A Framework for Single Source Cross-Domain Few-Shot Learning
Debasmit Das,Sungrack Yun,Fatih Porikli

Most current few-shot learning methods train a model from abundantly labeled base category data and then transfer and adapt the model to sparsely labeled novel category data. These methods mostly generalize well on novel categories from the same domain as the base categories but perform poorly for distant domain categories. In this paper, we propose a framework for few-shot learning coined as ConFeSS (Contrastive Learning and Feature Selection System) that tackles large domain shift between base and novel categories. The first step of our framework trains a feature extracting backbone with the contrastive loss on the base category data. Since the contrastive loss does not use supervision, the features can generalize better to distant target domains. For the second step, we train a masking module to select relevant features that are more suited to target domain classification. Finally, a classifier is fine-tuned along with the backbone such that the backbone produces features similar to the relevant ones. To evaluate our framework, we tested it on a recently introduced cross-domain few-shot learning benchmark. Experimental results demonstrate that our framework outperforms all meta-learning approaches and produces competitive results against recent cross-domain methods. Additional analyses are also performed to better understand our framework.
****

Revisiting and Advancing Fast Adversarial Training Through the lens of Bi-Level Optimization
Yihua Zhang,Guanhua Zhang,Prashant Khanduri,Mingyi Hong,Shiyu Chang,Sijia Liu

Adversarial training (AT) has become a widely recognized defense mechanism to improve the robustness of deep neural networks against adversarial attacks. It is originated from solving a min-max optimization problem, where the minimizer (i.e., defender) seeks a robust model to minimize the worst-case training loss at the presence of adversarial examples crafted by the maximizer (i.e., attacker). However,the min-max nature makes AT computationally intensive and thus difficult to scale. Thus, the problem of FAST-AT arises. Nearly all the recent progress is achieved based on the following simplification: The iterative attack generation method used in the maximization step of AT is replaced by the simplest one-shot gradient sign-based PGD method. Nevertheless, FAST-AT is far from satisfactory, and it lacks theoretically-grounded design. For example, a FAST-AT method may suffer from robustness catastrophic overfitting when training with strong adversaries.

In this paper, we foster a technological breakthrough for designing FAST-AT through the lens of bi-level optimization (BLO) instead of min-max optimization. First, we theoretically show that the most commonly-used algorithmic specification of FAST-AT is equivalent to the linearized BLO along the direction given by the sign of input gradient. Second, with the aid of BLO, we develop a new systematic and effective fast bi-level AT framework, termed FAST-BAT, whose algorithm is rigorously derived by leveraging the theory of implicit gradient. In contrast to FAST-AT, FAST-BAT has the least restriction to placing the tradeoff between computation efficiency and adversarial robustness. For example, it is capable of def ending sign-based projected gradient descent (PGD) attacks without calling any gradient sign method and explicit robust regularization during training. Furthermore, we empirically show that our method outperforms state-of-the-art FAST-AT baselines. In particular, FAST-BAT can achieve superior model robustness without inducing robustness catastrophic overfitting and losing standard accuracy.

****

Second-Order Rewards For Successor Features

Norman L Tasfi,Miriam Capretz

Current Reinforcement Learning algorithms have reached new heights in performance. However, such algorithms often require hundreds of millions of samples, often resulting in policies that are unable to transfer between tasks without full retraining. Successor features aim to improve this situation by decomposing the policy into two components: one capturing environmental dynamics and the other modelling reward. Where the reward function is formulated as the linear combination of learned state features and a learned parameter vector. Under this form, transfer between related tasks now only requires training the reward component. In this paper, we propose a novel extension to the successor feature framework resulting in a natural second-order variant. After derivation of the new state-action value function, a second additive term emerges, this term predicts reward as a non-linear combination of state features while providing additional benefits. Experimentally, we show that this term explicitly models the environment's stochasticity and can also be used in place of ϵ-greedy exploration methods during transfer. The performance of the proposed extension to the successor feature framework is validated empirically on a 2D navigation task, the control of a simulated robotic arm, and the Doom environment.

****

Granger causal inference on DAGs identifies genomic loci regulating transcription

Alexander P Wu,Rohit Singh,Bonnie Berger

When a dynamical system can be modeled as a sequence of observations, Granger causality is a powerful approach for detecting predictive interactions between its variables. However, traditional Granger causal inference has limited utility in domains where the dynamics need to be represented as directed acyclic graphs (DAGs) rather than as a linear sequence, such as with cell differentiation trajectories. Here, we present GrID-Net, a framework based on graph neural networks with lagged message passing for Granger causal inference on DAG-structured systems. Our motivating application is the analysis of single-cell multimodal data to identify genomic loci that mediate the regulation of specific genes. To our knowledge, GrID-Net is the first single-cell analysis tool that accounts for the temporal lag between a genomic locus becoming accessible and its downstream effect on a target gene's expression. We applied GrID-Net on multimodal single-cell assays that profile chromatin accessibility (ATAC-seq) and gene expression (RNA-seq) in the same cell and show that it dramatically outperforms existing methods for inferring regulatory locus-gene links, achieving up to 71% greater agreement with independent population genetics-based estimates. By extending Granger causality to DAG-structured dynamical systems, our work unlocks new domains for causal analyses and, more specifically, opens a path towards elucidating gene regulatory interactions relevant to cellular differentiation and complex human diseases at unprecedented scale and resolution.

****

Gating Mechanisms Underlying Sequence-to-Sequence Working Memory

Ian D Jordan,Piotr A Sokol,Il Memming Park

Working memory is the process by which a system temporarily stores information across a necessary duration. Memory retention and manipulation of discrete sequences are fundamental building blocks for the underlying computation required to perform working memory tasks. Recurrent neural networks (RNNs) have proven themselves to be powerful tools for such problems, as they, through training, bring rise to the dynamical behavior necessary to enact these computations over many time-steps. As of yet, the means by which these learned internal structures of the network result in a desired set of outputs remains broadly elusive. Furthermore, what is known is often difficult to extrapolate from due to a task specific formalism. In this work, we analyze an RNN, trained perfectly on a discrete sequence working memory task, in fine detail. We explain the learned mechanisms by which this network holds memory and extracts information from memory, and how gating is a natural architectural component to achieve these structures. A synthetic solution to a simplified variant of the working memory task is realized. We then explore how these results can be extrapolated to alternative tasks.

****

Randomized Primal-Dual Coordinate Method for Large-scale Linearly Constrained Nonsmooth Nonconvex Optimization

Lei Zhao,Daoli Zhu,Xiao Li

The large-scale linearly constrained nonsmooth nonconvex optimization finds wide applications in machine learning, including non-PSD Kernel SVM, linearly constrained Lasso with nonsmooth nonconvex penalty, etc. To tackle this class of optimization problems, we propose an efficient algorithm called Nonconvex Randomized Primal-Dual Coordinate (N-RPDC) method. At each iteration, this method only randomly selects a block of primal variables to update rather than updating all the variables, which is suitable for large-scale problems. We provide two types of convergence results for N-RPDC. We first show that any cluster point of the sequence of iterates generated by N-RPDC is almost surely (i.e., with probability 1) a stationary point. In addition, we also provide an almost sure asymptotic convergence rate of $O(1/\sqrt{k})$. Next, we establish the expected $O(\varepsilon^{-2})$ iteration complexity of N-RPDC in order to drive a natural stationarity measure below ε in expectation. The fundamental aspect to establishing the aforementioned convergence results is a \emph{surrogate stationarity measure} we discovered for analyzing N-RPDC. Finally, we conduct a set of experiments to show the efficacy of N-RPDC.

****

A Risk-Sensitive Policy Gradient Method

Jared Markowitz,Ryan Gardner,Ashley Llorens,Raman Arora,I-Jeng Wang

Standard deep reinforcement learning (DRL) agents aim to maximize expected reward, considering collected experiences equally in formulating a policy. This differs from human decision-making, where gains and losses are valued differently and outlying outcomes are given increased consideration. It also wastes an opportunity for the agent to modulate behavior based on distributional context. Several approaches to distributional DRL have been investigated, with one popular strategy being to evaluate the projected distribution of returns for possible actions. We propose a more direct approach, whereby the distribution of full-episode outcomes is optimized to maximize a chosen function of its cumulative distribution function (CDF). This technique allows for outcomes to be weighed based on relative quality, does not require modification of the reward function to modulate agent behavior, and may be used for both continuous and discrete action spaces. We show how to achieve an unbiased estimate of the policy gradient for a broad class of CDF-based objectives via sampling, subsequently incorporating variance reduction measures to facilitate effective on-policy learning. We use the resulting approach to train agents with different “risk profiles” in penalty-based formulations of six OpenAI Safety Gym environments, finding that moderate emphasis on improvement in training scenarios where the agent performs poorly generally improves agent behavior. We interpret and explore this observation, which leads to improved performance over the widely-used Proximal Policy Optimization algorithm in all environments tested.

****

Understanding Domain Randomization for Sim-to-real Transfer

Xiaoyu Chen,Jiachen Hu,Chi Jin,Lihong Li,Liwei Wang

Reinforcement learning encounters many challenges when applied directly in the real world. Sim-to-real transfer is widely used to transfer the knowledge learned from simulation to the real world. Domain randomization---one of the most popular algorithms for sim-to-real transfer---has been demonstrated to be effective in various tasks in robotics and autonomous driving. Despite its empirical successes, theoretical understanding on why this simple algorithm works is largely mi ssing. In this paper, we propose a theoretical framework for sim-to-real transfers, in which the simulator is modeled as a set of MDPs with tunable parameters (corresponding to unknown physical parameters such as friction). We provide sharp bounds on the sim-to-real gap---the difference between the value of policy returned by domain randomization and the value of an optimal policy for the real world. We prove that sim-to-real transfer can succeed under mild conditions without any real-world training samples. Our theory also highlights the importance of using memory (i.e., history-dependent policies) in domain randomization. Our proof is based on novel techniques that reduce the problem of bounding the sim-to-real gap to the problem of designing efficient learning algorithms for infinite-horizon MDPs, which we believe are of independent interest.

****

Energy-Inspired Molecular Conformation Optimization

Jiaqi Guan,Wesley Wei Qian,qiang liu,Wei-Ying Ma,Jianzhu Ma,Jian Peng

This paper studies an important problem in computational chemistry: predicting a molecule's spatial atom arrangements, or a molecular conformation. We propose a neural energy minimization formulation that casts the prediction problem into an unrolled optimization process, where a neural network is parametrized to learn the gradient fields of an implicit conformational energy landscape. Assuming different forms of the underlying potential energy function, we can not only reinterpret and unify many of the existing models but also derive new variants of SE(3)-equivariant neural networks in a principled manner. In our experiments, these new variants show superior performance in molecular conformation optimization comparing to existing SE(3)-equivariant neural networks. Moreover, our energy-inspired formulation is also suitable for molecular conformation generation, where we can generate more diverse and accurate conformers comparing to existing baselines.

****

Perturbation Deterioration: The Other Side of Catastrophic Overfitting

Zichao Li,Liyuan Liu,Chengyu Dong,Jingbo Shang

Our goal is to understand why the robustness accuracy would abruptly drop to zero, after conducting FGSM-style adversarial training for too long. While this phenomenon is commonly explained as overfitting, we observe that it is a twin process: not only does the model catastrophic overfits to one type of perturbation, but also the perturbation deteriorates into random noise. For example, at the same epoch when the FGSM-trained model catastrophically overfits, its generated perturbations deteriorate into random noise. Intuitively, once the generated perturbations become weak and inadequate, models would be misguided to overfit those weak attacks and fail to defend strong ones. In the light of our analyses, we propose APART, an adaptive adversarial training method, which parameterizes perturbation generation and progressively strengthens them. In our experiments, APART successfully prevents perturbation deterioration and catastrophic overfitting. Also, APART significantly improves the model robustness while maintaining the same efficiency as FGSM-style methods, e.g., on the CIFAR-10 dataset, APART achieves 53.89%accuracy under the PGD-20 attack and 49.05% accuracy under the AutoAttack.

****

Stability based Generalization Bounds for Exponential Family Langevin Dynamics

Arindam Banerjee,Tiancong Chen,Xinyan Li,Yingxue Zhou

We study the generalization of noisy stochastic mini-batch based iterative algorithms based on the notion of stability. Recent years have seen key advances in data-dependent generalization bounds for noisy iterative learning algorithms such as stochastic gradient Langevin dynamics (SGLD) based on (Mou et al., 2018; Li et al., 2020) and related approaches (Negrea et al., 2019; Haghifam et al., 2020). In this paper, we unify and substantially generalize stability based generalization bounds and make three technical advances. First, we bound the generalization error of general noisy stochastic iterative algorithms (not necessarily gradient descent) in terms of expected stability, which in turn can be bounded by the expected Le Cam Style Divergence (LSD). Such bounds have a $O(1/n)$ sample dependence unlike many existing bounds with $O(1/\sqrt{n})$ dependence. Second, we introduce Exponential Family Langevin Dynamics (EFLD) which is a substantial generalization of SGLD and which allows exponential family noise to be used with gradient descent. We establish data-dependent expected stability based generalization bounds for general EFLD. Third, we consider an important new special case of EFLD: Noisy Sign-SGD, which extends Sign-SGD by using Bernoulli noise over $\{-1,+1\}$, and we establish optimization guarantees for the algorithm. Further, we present empirical results on benchmark datasets to illustrate the our bounds are non-vacuous and quantitatively much sharper than existing bounds.

****

Towards Deepening Graph Neural Networks: A GNTK-based Optimization Perspective

Wei Huang,Yayong Li,weitao Du,Richard Xu,Jie Yin,Ling Chen,Miao Zhang

Graph convolutional networks (GCNs) and their variants have achieved great success in dealing with graph-structured data. Nevertheless, it is well known that deep GCNs suffer from the over-smoothing problem, where node representations tend to be indistinguishable as more layers are stacked up. The theoretical research to date on deep GCNs has focused primarily on expressive power rather than trainability, an optimization perspective. Compared to expressivity, trainability attempts to address a more fundamental question: Given a sufficiently expressive space of models, can we successfully find a good solution via gradient descent-based optimizers? This work fills this gap by exploiting the Graph Neural Tangent Kernel (GNTK), which governs the optimization trajectory under gradient descent for wide GCNs. We formulate the asymptotic behaviors of GNTK in the large depth, which enables us to reveal the dropping trainability of wide and deep GCNs at an exponential rate in the optimization process. Additionally, we extend our theoretical framework to analyze residual connection-based techniques, which are found to be merely able to mitigate the exponential decay of trainability mildly. Inspired by our theoretical insights on trainability, we propose Critical DropEdge, a connectivity-aware and graph-adaptive sampling method, to alleviate the exponential decay problem more fundamentally. Experimental evaluation consistently confirms using our proposed method can achieve better results compared to relevant counterparts with both infinite-width and finite-width.

****

Meta-free few-shot learning via representation learning with weight averaging

Kuilin Chen,Chi-Guhn Lee

Recent studies on few-shot classification using transfer learning pose challenges to the effectiveness and efficiency of episodic meta-learning algorithms. Transfer learning approaches are a natural alternative, but they are restricted to few-shot classification. Moreover, little attention has been on the development of probabilistic models with well-calibrated uncertainty from few-shot samples, except for some Bayesian episodic learning algorithms. To tackle the aforementioned issues, we propose a new transfer learning method to obtain accurate and reliable models for few-shot regression and classification. The resulting method does not require episodic meta-learning and is called meta-free representation learning (MFRL). MFRL first finds low-rank representation generalizing well on meta-test tasks. Given the learned representation, probabilistic linear models are fine-tuned with few-shot samples to obtain models with well-calibrated uncertainty. The proposed method not only achieves the highest accuracy on a wide range of few-shot learning benchmark datasets but also correctly quantifies the prediction uncertainty. In addition, weight averaging and temperature scaling are effective in improving the accuracy and reliability of few-shot learning in existing meta-learning algorithms with a wide range of learning paradigms and model architectures.

****

Connectome-constrained Latent Variable Model of Whole-Brain Neural Activity

Lu Mi,Richard Xu,Sridhama Prakhya,Albert Lin,Nir Shavit,Aravinthan Samuel,Srinivas C Turaga

The availability of both anatomical connectivity and brain-wide neural activity measurements in C. elegans make the worm a promising system for learning detailed, mechanistic models of an entire nervous system in a data-driven way. However, one faces several challenges when constructing such a model. We often do not ha ve direct experimental access to important modeling details such as single-neuron dynamics and the signs and strengths of the synaptic connectivity. Further, neural activity can only be measured in a subset of neurons, often indirectly via calcium imaging, and significant trial-to-trial variability has been observed. To address these challenges, we introduce a connectome-constrained latent variable model (CC-LVM) of the unobserved voltage dynamics of the entire C. elegans nervous system and the observed calcium signals. We used the framework of variational autoencoders to fit parameters of the mechanistic simulation constituting the generative model of the LVM to calcium imaging observations. A variational approximate posterior distribution over latent voltage traces for all neurons is efficiently inferred using an inference network, and constrained by a prior distribution given by the biophysical simulation of neural dynamics. We applied this model to an experimental whole-brain dataset, and found that connectomic constraints enable our LVM to predict the activity of neurons whose activity were withheld significantly better than models unconstrained by a connectome. We explored models with different degrees of biophysical detail, and found that models with realistic conductance-based synapses provide markedly better predictions than current-based synapses for this system.

****

NAS-Bench-360: Benchmarking Diverse Tasks for Neural Architecture Search

Renbo Tu,Mikhail Khodak,Nicholas Carl Roberts,Ameet Talwalkar

Most existing neural architecture search (NAS) benchmarks and algorithms prioritize performance on well-studied tasks, e.g., image classification on CIFAR and ImageNet. This makes the applicability of NAS approaches in more diverse areas inadequately understood.

In this paper, we present NAS-Bench-360, a benchmark suite for evaluating state-of-the-art NAS methods for convolutional neural networks (CNNs). To construct it, we curate a collection of ten tasks spanning a diverse array of application domains, dataset sizes, problem dimensionalities, and learning objectives. By carefully selecting tasks that can both interoperate with modern CNN-based search methods but that are also far-afield from their original development domain, we can use NAS-Bench-360 to investigate the following central question: do existing state-of-the-art NAS methods perform well on diverse tasks? Our experiments show that a modern NAS procedure designed for image classification can indeed find good architectures for tasks with other dimensionalities and learning objectives; however, the same method struggles against more task-specific methods and performs catastrophically poorly on classification in non-vision domains. The case for NAS robustness becomes even more dire in a resource-constrained setting, where a recent NAS method provides little-to-no benefit over much simpler baselines. These results demonstrate the need for a benchmark such as NAS-Bench-360 to help develop NAS approaches that work well on a variety of tasks, a crucial component of a truly robust and automated pipeline. We conclude with a demonstration of the kind of future research our suite of tasks will enable. All data and code is made publicly available.

****

T-WaveNet: A Tree-Structured Wavelet Neural Network for Time Series Signal Analysis

Minhao LIU,Ailing Zeng,Qiuxia LAI,Ruiyuan Gao,Min Li,Jing Qin,Qiang Xu

Time series signal analysis plays an essential role in many applications, e.g., activity recognition and healthcare monitoring.

Recently, features extracted with deep neural networks (DNNs) have shown to be more effective than conventional hand-crafted ones.

However, most existing solutions rely solely on the network to extract information carried in the raw signal, regardless of its inherent physical and statistical properties, leading to sub-optimal performance particularly under a limited amount of training data.

In this work, we propose a novel tree-structured wavelet neural network for time series signal analysis, namely \emph{T-WaveNet}, taking advantage of an inherent property of various types of signals, known as the \emph{dominant frequency range}. Specifically, with \emph{T-WaveNet}, we first conduct frequency spectrum e

nergy analysis of the signals to get a set of dominant frequency subbands. Then, we construct a tree-structured network that iteratively decomposes the input signal into various frequency subbands with similar energies. Each node on the tree is built with an invertible neural network (INN) based wavelet transform unit. Such a disentangled representation learning method facilitates a more effective extraction of the discriminative features, as demonstrated with the comprehensive experiments on various real-life time series classification datasets.
****

An Interpretable Graph Generative Model with Heterophily
Sudhanshu Chanpuriya,Ryan Rossi,Anup Rao,Tung Mai,Nedim Lipka,Zhao Song,Cameron N Musco
Many models for graphs fall under the framework of edge-independent dot product models. These models output the probabilities of edges existing between all pairs of nodes, and the probability of a link between two nodes increases with the dot product of vectors associated with the nodes. Recent work has shown that these models are unable to capture key structures in real-world graphs, particularly heterophilous structures, wherein links occur between dissimilar nodes. We propose the first edge-independent graph generative model that is a) expressive enough to capture heterophily, b) produces nonnegative embeddings, which allow link predictions to be interpreted in terms of communities, and c) optimizes effectively on real-world graphs with gradient descent on a cross-entropy loss. Our theoretical results demonstrate the expressiveness of our model in its ability to exactly reconstruct a graph using a number of clusters that is linear in the maximum degree, along with its ability to capture both heterophily and homophily in the data. Further, our experiments demonstrate the effectiveness of our model for a variety of important application tasks such as multi-label clustering and link prediction.
****

$\mathrm{SO}(2)$-Equivariant Reinforcement Learning
Dian Wang,Robin Walters,Robert Platt
Equivariant neural networks enforce symmetry within the structure of their convolutional layers, resulting in a substantial improvement in sample efficiency when learning an equivariant or invariant function. Such models are applicable to robotic manipulation learning which can often be formulated as a rotationally symmetric problem. This paper studies equivariant model architectures in the context of Q-learning and actor-critic reinforcement learning. We identify equivariant and invariant characteristics of the optimal Q-function and the optimal policy and propose equivariant DQN and SAC algorithms that leverage this structure. We present experiments that demonstrate that our equivariant versions of DQN and SAC can be significantly more sample efficient than competing algorithms on an important class of robotic manipulation problems.
****

Test Time Robustification of Deep Models via Adaptation and Augmentation
Marvin Mengxin Zhang,Sergey Levine,Chelsea Finn
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 demonstrate that this approach consistently improves robust ResNet and vision transformer models, achieving accuracy gains of 1-8% over standard model evaluation and also generally outperforming prior augmentation and adaptation strategies. We achieve state-of-the-art results for test shifts caused by image corruptions (ImageNet-C), renditions of common objects (ImageNet-R), and, among ResNet-50 models, adversarially chosen natural examples (ImageNet-A).
****

ACTIVE REFINEMENT OF WEAKLY SUPERVISED MODELS
Mononito Goswami,Chufan Gao,Benedikt Boecking,Saswati Ray,Artur Dubrawski

Supervised machine learning (ML) has fueled major advances in several domains such as health, education and governance. However, most modern ML methods rely on vast quantities of point-by-point hand-labeled training data. In domains such as clinical research, where data collection and its careful characterization is particularly expensive and tedious, this reliance on pointillisticaly labeled data is one of the biggest roadblocks to the adoption of modern data-hungry ML algorithms. Data programming, a framework for learning from weak supervision, attempts to overcome this bottleneck by generating probabilistic training labels from simple yet imperfect heuristics obtained a priori from domain experts. We present WARM, Active Refinement of Weakly Supervised Models, a principled approach to iterative and interactive improvement of weakly supervised models via active learning. WARM directs domain experts' attention on a few selected data points that, when annotated, would improve the label model's probabilistic output in terms of accuracy the most. Gradient backpropagation is then used to iteratively update decision parameters of the heuristics of the label model. Experiments on multiple real-world medical classification datasets reveal that WARM can substantially improve the accuracy of probabilistic labels, a direct measure of training data quality, with as few as 30 queries to clinicians. Additional experiments with domain shift and artificial noise in the LFs, demonstrate WARM's ability to adapt heuristics and the end model to changing population characteristics as well as its robustness to mis-specification of domain-expert-acquired LFs. These capabilities make WARM a potentially useful tool for deploying, maintaining, and auditing weakly supervised systems in practice.
****

SemiRetro: Semi-template framework boosts deep retrosynthesis prediction
Zhangyang Gao,Cheng Tan,Lirong Wu,Haitao Lin,Stan Z. Li

Retrosynthesis brings scientific and societal benefits by inferring possible reaction routes toward novel molecules. Recently, template-based (TB) and template-free (TF) molecule graph learning methods have shown promising results to solve this problem. TB methods are more accurate using pre-encoded reaction templates, and TF methods are more scalable by decomposing retrosynthesis into subproblems, i.e., center identification and synthon completion. To combine both advantages of TB and TF, we suggest breaking a full-template into several semi-templates and embedding them into the two-step TF framework. Since many semi-templates are reduplicative, the template redundancy can be reduced while the essential chemical knowledge is still preserved to facilitate synthon completion. We call our method SemiRetro and introduce a directed relational graph attention (DRGAT) layer to extract expressive features for better center identification. Experimental results show that SemiRetro significantly outperforms both existing TB and TF methods. In scalability, SemiRetro covers 96.9\% data using 150 semi-templates, while previous template-based GLN requires 11,647 templates to cover 93.3\% data. In top-1 accuracy, SemiRetro exceeds template-free G2G 3.4\% (class known) and 6.4\% (class unknown). Besides, SemiReto has better interpretability and training efficiency than existing methods.
****

Scarf: Self-Supervised Contrastive Learning using Random Feature Corruption
Dara Bahri,Heinrich Jiang,Yi Tay,Donald Metzler

Self-supervised contrastive representation learning has proved incredibly successful in the vision and natural language domains, enabling state-of-the-art performance with orders of magnitude less labeled data. However, such methods are domain-specific and little has been done to leverage this technique on real-world \

emph{tabular} datasets. We propose \textsc{Scarf}, a simple, widely-applicable technique for contrastive learning, where views are formed by corrupting a random subset of features. When applied to pre-train deep neural networks on the 69 real-world, tabular classification datasets from the OpenML-CC18 benchmark, \textsc{Scarf} not only improves classification accuracy in the fully-supervised setting but does so also in the presence of label noise and in the semi-supervised setting where only a fraction of the available training data is labeled. We show that \textsc{Scarf} complements existing strategies and outperforms alternatives like autoencoders. We conduct comprehensive ablations, detailing the importance of a range of factors.

****

Trans-Encoder: Unsupervised sentence-pair modelling through self- and mutual-distillations

Fangyu Liu,Yunlong Jiao,Jordan Massiah,Emine Yilmaz,Serhii Havrylov

In NLP, a large volume of tasks involve pairwise comparison between two sequences (e.g. sentence similarity and paraphrase identification). Predominantly, two formulations are used for sentence-pair tasks: bi-encoders and cross-encoders. Bi-encoders produce fixed-dimensional sentence representations and are computationally efficient, however, they usually underperform cross-encoders. Cross-encoders can leverage their attention heads to exploit inter-sentence interactions for better performance but they require task fine-tuning and are computationally more expensive. In this paper, we present a completely unsupervised sentence representation model termed as Trans-Encoder that combines the two learning paradigms into an iterative joint framework to simultaneously learn enhanced bi- and cross-encoders. Specifically, on top of a pre-trained Language Model (PLM), we start with converting it to an unsupervised bi-encoder, and then alternate between the bi- and cross-encoder task formulations. In each alternation, one task formulation will produce pseudo-labels which are used as learning signals for the other task formulation. We then propose an extension to conduct such self-distillation approach on multiple PLMs in parallel and use the average of their pseudo-labels for mutual distillation. Trans-Encoder creates, to the best of our knowledge, the first completely unsupervised cross-encoder and also a state-of-the-art unsupervised bi-encoder for sentence similarity. Both the bi-encoder and cross-encoder formulations of Trans-Encoder outperform recently proposed state-of-the-art unsupervised sentence encoders such as Mirror-BERT and SimCSE by up to 5% on the sentence similarity benchmarks.

****

Path Integral Sampler: A Stochastic Control Approach For Sampling

Qinsheng Zhang,Yongxin Chen

We present Path Integral Sampler~(PIS), a novel algorithm to draw samples from unnormalized probability density functions. The PIS is built on the Schr\"odinger bridge problem which aims to recover the most likely evolution of a diffusion process given its initial distribution and terminal distribution. The PIS draws samples from the initial distribution and then propagates the samples through the Schr\"odinger bridge to reach the terminal distribution. Applying the Girsanov theorem, with a simple prior diffusion, we formulate the PIS as a stochastic optimal control problem whose running cost is the control energy and terminal cost is chosen according to the target distribution. By modeling the control as a neural network, we establish a sampling algorithm that can be trained end-to-end. We provide theoretical justification of the sampling quality of PIS in terms of Wasserstein distance when sub-optimal control is used. Moreover, the path integrals theory is used to compute importance weights of the samples to compensate for the bias induced by the sub-optimality of the controller and the time-discretization. We experimentally demonstrate the advantages of PIS compared with other start-of-the-art sampling methods on a variety of tasks.

****

Style Equalization: Unsupervised Learning of Controllable Generative Sequence Models

Jen-Hao Rick Chang,Ashish Shrivastava,Hema Swetha Koppula,Xiaoshuai Zhang,Oncel Tuzel

Controllable generative sequence models with the capability to extract and replicate the style of specific examples enable many applications, including narrating audiobooks in different voices, auto-completing and auto-correcting written handwriting, and generating missing training samples for downstream recognition tasks. However, typical training algorithms for these controllable sequence generative models suffer from the training-inference mismatch, where the same sample is used as content and style input during training but different samples are given during inference. In this paper, we tackle the training-inference mismatch encountered during unsupervised learning of controllable generative sequence models. By introducing a style transformation module that we call style equalization, we enable training using different content and style samples and thereby mitigate the training- inference mismatch. To demonstrate its generality, we applied style equalization to text-to-speech and text-to-handwriting synthesis on three datasets. Our models achieve state-of-the-art style replication with a similar mean style opinion score as the real data. Moreover, the proposed method enables style interpolation between sequences and generates novel styles.

****

Responsible Disclosure of Generative Models Using Scalable Fingerprinting

Ning Yu,Vladislav Skripniuk,Dingfan Chen,Larry S. Davis,Mario Fritz

Over the past years, deep generative models have achieved a new level of performance. Generated data has become difficult, if not impossible, to be distinguished from real data. While there are plenty of use cases that benefit from this technology, there are also strong concerns on how this new technology can be misused to generate deep fakes and enable misinformation at scale. Unfortunately, current deep fake detection methods are not sustainable, as the gap between real and fake continues to close. In contrast, our work enables a responsible disclosure of such state-of-the-art generative models, that allows model inventors to fingerprint their models, so that the generated samples containing a fingerprint can be accurately detected and attributed to a source. Our technique achieves this by an efficient and scalable ad-hoc generation of a large population of models with distinct fingerprints. Our recommended operation point uses a 128-bit fingerprint which in principle results in more than 10^{38} identifiable models. Experiments show that our method fulfills key properties of a fingerprinting mechanism and achieves effectiveness in deep fake detection and attribution. Code and models are available at https://github.com/ningyu1991/ScalableGANFingerprints.

****

Model Zoo: A Growing Brain That Learns Continually

Rahul Ramesh,Pratik Chaudhari

This paper argues that continual learning methods can benefit by splitting the capacity of the learner across multiple models. We use statistical learning theory and experimental analysis to show how multiple tasks can interact with each other in a non-trivial fashion when a single model is trained on them. The generalization error on a particular task can improve when it is trained with synergistic tasks, but can also deteriorate when trained with competing tasks. This theory motivates our method named Model Zoo which, inspired from the boosting literature, grows an ensemble of small models, each of which is trained during one episode of continual learning. We demonstrate that Model Zoo obtains large gains in accuracy on a wide variety of continual learning benchmark problems.

****

DEUP: Direct Epistemic Uncertainty Prediction

Moksh Jain,Salem Lahlou,Hadi Nekoei,Victor I Butoi,Paul Bertin,Jarrid Rector-Brooks,Maksym Korablyov,Yoshua Bengio

Epistemic uncertainty is the part of out-of-sample prediction error due to the lack of knowledge of the learner. Whereas previous work was focusing on model variance, we propose a principled approach for directly estimating epistemic uncertainty by learning to predict generalization error and subtracting an estimate of aleatoric uncertainty, i.e., intrinsic unpredictability. This estimator of epistemic uncertainty includes the effect of model bias (or misspecification) and is useful in interactive learning environments arising in active learning or reinforcement learning. In addition to discussing these properties of Direct Epistemi c Uncertainty Prediction (DEUP), we illustrate its advantage against existing methods for uncertainty estimation on downstream tasks including sequential model optimization and reinforcement learning. We also evaluate the quality of uncertainty estimates from DEUP for probabilistic classification of images and for estimating uncertainty about synergistic drug combinations.

****

Why be adversarial? Let's cooperate!: Cooperative Dataset Alignment via JSD Upper Bound

Wonwoong Cho,Ziyu Gong,David I. Inouye

Unsupervised dataset alignment estimates a transformation that maps two or more source domains to a shared aligned domain given only the domain datasets.

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.

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 dataset alignment.

We present empirical results of our framework on both simulated and real-world datasets to demonstrate the benefits of our approach.

****

Image Compression and Classification Using Qubits and Quantum Deep Learning

Ali Mohsen,Mo Tiwari

Recent work suggests that quantum machine learning techniques can be used for classical image classification by encoding the images in quantum states and using a quantum neural network for inference. However, such work has been restricted to very small input images, at most 4×4, that are unrealistic and cannot even be accurately labeled by humans. The primary difficulties in using larger input images is that hitherto-proposed encoding schemes necessitate more qubits than are physically realizable. We propose a framework to classify larger, realistic images using quantum systems. Our approach relies on a novel encoding mechanism that embeds images in quantum states while necessitating fewer qubits than prior work. Our framework is able to classify images that are larger than previously possible, up to 16×16 for the MNIST dataset on a personal laptop, and obtains accuracy comparable to classical neural networks with the same number of learnable parameters. We also propose a technique for further reducing the number of qubits needed to represent images that may result in an easier physical implementation at the expense of final performance. Our work enables quantum machine learning and classification on classical datasets of dimensions that were previously intractable by physically realizable quantum computers or classical simulation.

****

Fully Online Meta-Learning Without Task Boundaries

Jathushan Rajasegaran,Chelsea Finn,Sergey Levine

While deep networks can learn complex classifiers and models, many applications require models that continually adapt to changing input distributions, changing tasks, and changing environmental conditions. Indeed, this ability to continuously accrue knowledge and use the past experience to learn new tasks quickly in continual settings is one of the key properties of an intelligent system. For complex and high-dimensional problems, simply updating the model continually with standard learning algorithms such as gradient descent may result in slow adaptation. Meta-learning can provide a powerful tool to accelerate adaptation but is conventionally studied in batch settings. In this paper, we study how meta-learning can be applied to tackle online problems of this nature, simultaneously adapting to online to changing tasks and input distributions and meta-training the mode l in order to adapt more quickly in the future. Extending meta-learning into the online setting presents its own challenges, and although several prior methods have studied related problems, they generally require a discrete notion of tasks, with known ground-truth task boundaries. Such methods typically adapt to each task in sequence, resetting the model between tasks, rather than adapting continuously across tasks. In many real-world settings, such discrete boundaries are unavailable, and may not even exist. To address these settings, we propose a Fully Online Meta-Learning (FOML) algorithm, which does not require any ground truth knowledge about the task boundaries and stays fully online without resetting back to pre-trained weights. Our experiments show that FOML was able to learn new tasks faster than the state-of-the-art online learning methods on Rainbow-MNIST, and CIFAR100 datasets.

****

Predicting Physics in Mesh-reduced Space with Temporal Attention

XU HAN,Han Gao,Tobias Pfaff,Jian-Xun Wang,Liping Liu

Auto-regressive sequence models for physics prediction are often restricted to low-dimensional systems, as memory cost increases with both spatial extents and sequence length. On the other hand, graph-based next-step prediction models have recently been very successful in modeling complex high-dimensional physical systems on irregular meshes, but suffer from error accumulation and drift, due to their short temporal attention span. In this paper, we present a method that marries the strengths of both approaches. We use a GNN to locally summarize features and create coarsened, compact mesh representation of the system state, onto which we apply a transformer-style temporal attention module. We use a second GNN to decode these predictions back to a full-sized graph and perform fine-scale updates. Our method outperforms a competitive GNN baseline on three complex fluid dynamics prediction tasks, from sonic shocks to vascular flow. We demonstrate stable rollouts without the need for training noise and show perfectly phase-stable predictions even for very long sequences. More broadly, we believe our approach paves the way to bringing the benefits of attention-based sequence models to solving high-dimensional complex physics tasks.

****

Weakly Supervised Label Learning Flows

You Lu,Chidubem Gibson Arachie,Bert Huang

Supervised learning usually requires a large amount of labelled data. However, attaining ground-truth labels is costly for many tasks. Alternatively, weakly supervised methods learn with only cheap weak signals that only approximately label some data. Many existing weakly supervised learning methods learn a deterministic function that estimates labels given the input data and weak signals. In this paper, we develop label learning flow (LLF), a general framework for weakly supervised learning problems. Our method is a generative model based on normalizing flows. The main idea of LLF is to optimize the conditional likelihoods of all possible labelings of the data within a constrained space defined by weak signals. We develop a training method for LLF that trains the conditional flow inversely and avoids estimating the labels. Once a model is trained, we can make predictions with a sampling algorithm. We apply LLF to three weakly supervised learning problems. Experiment results show that our method outperforms many state-of-the-art alternatives.

****

Divergent representations of ethological visual inputs emerge from supervised, unsupervised, and reinforcement learning

Grace W Lindsay,Josh Merel,Thomas D. Mrsic-Flogel,Maneesh Sahani

Artificial neural systems trained using reinforcement, supervised, and unsupervised learning all acquire internal representations of high dimensional input. To what extent these representations depend on the different learning objectives is largely unknown. Here we compare the representations learned by eight different convolutional neural networks, each with identical ResNet architectures and trained on the same family of egocentric images, but embedded within different learning systems. Specifically, the representations are trained to guide action in a compound reinforcement learning task; to predict one or a combination of three task-related targets with supervision; or using one of three different unsupervised objectives. Using representational similarity analysis, we find that the network trained with reinforcement learning differs most from the other networks. Through further analysis using metrics inspired by the neuroscience literature, we find that the model trained with reinforcement learning has a high-dimensional representation wherein individual images are represented with very different patterns of neural activity. These representations seem to arise in order to guide long-term behavior and goal-seeking in the RL agent. Our results provide insights into how the properties of neural representations are influenced by objective functions and can inform transfer learning approaches.

****

How unlabeled data improve generalization in self-training? A one-hidden-layer theoretical analysis

Shuai Zhang,Meng Wang,Sijia Liu,Pin-Yu Chen,Jinjun Xiong

Self-training, a semi-supervised learning algorithm, leverages a large amount of unlabeled data to improve learning when the labeled data are limited. Despite empirical successes, its theoretical characterization remains elusive. To the best of our knowledge, this work establishes the first theoretical analysis for the known iterative self-training paradigm and formally proves the benefits of unlabeled data in both training convergence and generalization ability. To make our theoretical analysis feasible, we focus on the case of one-hidden-layer neural networks. However, theoretical understanding of iterative self-training is non-trivial even for a shallow neural network. One of the key challenges is that existing neural network landscape analysis built upon supervised learning no longer holds in the (semi-supervised) self-training paradigm. We address this challenge and prove that iterative self-training converges linearly with both convergence rate and generalization accuracy improved in the order of $1/\sqrt{M}$, where M is the number of unlabeled samples. Extensive experiments from shallow neural networks to deep neural networks are also provided to justify the correctness of our established theoretical insights on self-training.

****

SABAL: Sparse Approximation-based Batch Active Learning

Maohao Shen,Bowen Jiang,Jacky Y. Zhang,Oluwasanmi O Koyejo

We propose a novel and general framework (i.e., SABAL) that formulates batch active learning as a sparse approximation problem. SABAL aims to find a weighted subset from the unlabeled data pool such that the corresponding training loss function approximates its full data pool counterpart. We realize the general framework as a sparsity-constrained discontinuous optimization problem that explicitly balances uncertainty and representation for large-scale applications, for which we propose both greedy and iterative hard thresholding schemes. The proposed method can adapt to various settings, including both Bayesian and non-Bayesian neural networks. Numerical experiments show that that SABAL achieves state-of-the-art performance across different settings with lower computational complexity.

****

Learning to Dequantise with Truncated Flows

Shawn Tan,Chin-Wei Huang,Alessandro Sordoni,Aaron Courville

Dequantisation is a general technique used for transforming data described by a discrete random variable x into a continuous (latent) random variable z, for the purpose of it being modeled by likelihood-based density models. Dequantisation was first introduced in the context of ordinal data, such as image pixel values. However, when the data is categorical, the dequantisation scheme is not obvious.

We learn such a dequantisation scheme $q(z | x)$, using variational inference with TRUncated FLows (TRUFL) --- a novel flow-based model that allows the dequantiser to have a learnable truncated support. Unlike previous work, the TRUFL dequantiser is (i) capable of embedding the data losslessly in certain cases, since the truncation allows the conditional distributions $q(z | x)$ to have non-overlapping bounded supports, while being (ii) trainable with back-propagation. Addtionally, since the support of the marginal $q(z)$ is bounded and the support of prior $p(z)$ is not, we propose renormalising the prior distribution over the supp ort of $q(z)$. We derive a lower bound for training, and propose a rejection sampling scheme to account for the invalid samples during generation. Experimentally, we benchmark TRUFL on constrained generation tasks, and find that it outperforms prior approaches. In addition, we find that rejection sampling results in higher validity for the constrained problems.

****

Permutation invariant graph-to-sequence model for template-free retrosynthesis and reaction prediction

Zhengkai Tu,Connor W. Coley

Synthesis planning and reaction outcome prediction are two fundamental problems in computer-aided organic chemistry for which a variety of data-driven approaches have emerged. Natural language approaches that model each problem as a SMILES-to-SMILES translation lead to a simple end-to-end formulation, reduce the need for data preprocessing, and enable the use of well-optimized machine translation model architectures. However, SMILES representations are not an efficient representation for capturing information about molecular structure, as evidenced by the success of SMILES augmentation to boost empirical performance. Here, we describe a novel Graph2SMILES model that combines the power of Transformer models for text generation with the permutation invariance of molecular graph encoders. As an end-to-end architecture, Graph2SMILES can be used as a drop-in replacement for the Transformer in any task involving molecule(s)-to-molecule(s) transformations. In our encoder, an attention-augmented directed message passing neural network (D-MPNN) captures local chemical environments, and the global attention encoder allows for long-range and intermolecular interactions, enhanced by graph-aware positional embedding. Graph2SMILES improves the top-1 accuracy of the Transformer baselines by 1.7% and 1.9% for reaction outcome prediction on USPTO_480k and USPTO_STEREO datasets respectively, and by 9.8% for one-step retrosynthesis on the USPTO_50k dataset.

****

Curriculum learning as a tool to uncover learning principles in the brain

Daniel R. Kepple,Rainer Engelken,Kanaka Rajan

We present a novel approach to use curricula to identify principles by which a system learns. Previous work in curriculum learning has focused on how curricula can be designed to improve learning of a model on particular tasks. We consider the inverse problem: what can a curriculum tell us about how a learning system acquired a task? Using recurrent neural networks (RNNs) and models of common experimental neuroscience tasks, we demonstrate that curricula can be used to differentiate learning principles using target-based and a representation-based loss functions as use cases. In particular, we compare the performance of RNNs using target-based learning rules versus those using representational learning rules on three different curricula in the context of two tasks. We show that the learned state-space trajectories of RNNs trained by these two learning rules under all curricula tested are indistinguishable. However, by comparing learning times during different curricula, we can disambiguate the learning rules and challenge traditional approaches of interrogating learning systems. Although all animals in neuroscience lab settings are trained by curriculum-based procedures called shaping, almost no behavioral or neural data are collected or published on the relative successes or training times under different curricula. Our results motivate the systematic collection and curation of data during shaping by demonstrating curriculum learning in RNNs as a tool to probe and differentiate learning principles used by biological systems, over conventional statistical analyses of learned state spaces.

****

Intra-class Mixup for Out-of-Distribution Detection

Deepak Ravikumar,Sangamesh Kodge,Isha Garg,Kaushik Roy

Deep neural networks have found widespread adoption in solving image recognition and natural language processing tasks. However, they make confident mispredictions when presented with data that does not belong to the training distribution, i.e. out-of-distribution (OoD) samples. Inter-class mixup has been shown to improve model calibration aiding OoD detection. However, we show that both empirical risk minimization and inter-class mixup create large angular spread in latent representation. This reduces the separability of in-distribution data from OoD data. In this paper we propose intra-class mixup supplemented with angular margin to improve OoD detection. Angular margin is the angle between the decision boundary normal and sample representation. We show that intra-class mixup forces the network to learn representations with low angular spread in the latent space. This improves the separability of OoD from in-distribution examples. Our approach when applied to various existing OoD detection techniques shows an improvement of 4.68% and 6.38% in AUROC performance over empirical risk minimization and inter-class mixup, respectively. Further, our approach aided with angular margin improves AUROC performance by 7.36% and 9.10% over empirical risk minimization and inter-class mixup, respectively.

****

Explaining Off-Policy Actor-Critic From A Bias-Variance Perspective

Ting-Han Fan,Peter Ramadge

Off-policy Actor-Critic algorithms have demonstrated phenomenal experimental performance but still require better explanations. To this end, we show its policy evaluation error on the distribution of transitions decomposes into: a Bellman error, a bias from policy mismatch, and a variance term from sampling. By comparing the magnitude of bias and variance, we explain the success of the Emphasizing Recent Experience sampling and 1/age weighted sampling. Both sampling strategies yield smaller bias and variance and are hence preferable to uniform sampling.

****

Neural Circuit Architectural Priors for Embodied Control

Nikhil Xie Bhattasali,Anthony M. Zador,Tatiana A Engel

Artificial neural networks coupled with learning-based methods have enabled robots to tackle increasingly complex tasks, but often at the expense of requiring large amounts of learning experience. In nature, animals are born with highly structured connectivity in their brains and nervous systems that enables them to efficiently learn robust motor skills. Capturing some of this structure in artificial models may bring robots closer to matching animal performance and efficiency. In this paper, we present Neural Circuit Architectural Priors (NCAP), a set of reusable architectural components and design principles for deriving network architectures for embodied control from biological neural circuits. We apply this method to control a simulated agent performing a locomotion task and show that the NCAP architecture achieves comparable asymptotic performance with fully connected MLP architectures while dramatically improving data efficiency and requiring far fewer parameters. We further show through an ablation analysis that principled excitation/inhibition and initialization play significant roles in our NCAP architecture. Overall, our work suggests a way of advancing artificial intelligence and robotics research inspired by systems neuroscience.

****

Optimizer Amalgamation

Tianshu Huang,Tianlong Chen,Sijia Liu,Shiyu Chang,Lisa Amini,Zhangyang Wang

Selecting an appropriate optimizer for a given problem is of major interest for researchers and practitioners. Many analytical optimizers have been proposed using a variety of theoretical and empirical approaches; however, none can offer a universal advantage over other competitive optimizers. We are thus motivated to study a new problem named Optimizer Amalgamation: how can we best combine a pool of "teacher" optimizers into a single "student" optimizer that can have stronger problem-specific performance? In this paper, we draw inspiration from the field of "learning to optimize" to use a learnable amalgamation target. First, we define three differentiable amalgamation mechanisms to amalgamate a pool of analytical optimizers by gradient descent. Then, in order to reduce variance of the amalgamation process, we also explore methods to stabilize the amalgamation process by perturbing the amalgamation target. Finally, we present experiments showing the superiority of our amalgamated optimizer compared to its amalgamated components and learning to optimize baselines, and the efficacy of our variance reducing perturbations.

****
DEEP GRAPH TREE NETWORKS
Nan Wu,Chaofan Wang

We propose Graph Tree Networks (GTree), a self-interpretive deep graph neural network architecture which originates from the tree representation of the graphs. In the tree representation, each node forms its own tree where the node itself is the root node and all its neighbors up to hop-k are the subnodes. Under the tree representation, the message propagates upward from the leaf nodes to the root node naturally and straightforwardly to update the root node's hidden features. This message passing (or neighborhood aggregation) scheme is essentially different from that in the vanilla GCN, GAT and many of their derivatives, and is demonstrated experimentally a superior message passing scheme. Models adopting this scheme has the capability of going deep. Two scalable graph learning models are proposed within this GTree network architecture - Graph Tree Convolution Network (GTCN) and Graph Tree Attention Network (GTAN), with demonstrated state-of-the-art performances on several benchmark datasets. The deep capability is also demonstrated for both models.

****
On the Impact of Hard Adversarial Instances on Overfitting in Adversarial Training
Chen Liu,Zhichao Huang,Mathieu Salzmann,Tong Zhang,Sabine Süsstrunk

Adversarial training is a popular method to robustify models against adversarial attacks.
However, it exhibits much more severe overfitting than training on clean inputs.
In this work, we investigate this phenomenon from the perspective of training instances, i.e., training input-target pairs.
To this end, we provide a quantitative and model-agnostic metric measuring the difficulty of an instance in the training set and analyze the model's behavior on instances of different difficulty levels.
This lets us show that the decay in generalization performance of adversarial training is a result of the model's attempt to fit hard adversarial instances.
We theoretically verify our observations for both linear and general nonlinear models, proving that models trained on hard instances have worse generalization performance than ones trained on easy instances.
In addition, this gap in generalization performance is larger in adversarial training.
Finally, we investigate solutions to mitigating adversarial overfitting in several scenarios, including when relying on fast adversarial training and in the context of fine-tuning a pretrained model with additional data.
Our results demonstrate adaptively using training data can improve model's robustness.

****
How to decay your learning rate
Aitor Lewkowycz

Complex learning rate schedules have become an integral part of deep learning. We find empirically that common fine-tuned schedules decay the learning rate after the weight norm bounces. This leads to the proposal of ABEL: an automatic scheduler which decays the learning rate by keeping track of the weight norm. ABEL's performance matches that of tuned schedules, is more robust with respect to its parameters and does not depend on the time budget. Through extensive experiments in vision, NLP, and RL, we show that if the weight norm does not bounce, we can simplify schedules even further with no loss in performance. In such cases, a complex schedule has similar performance to a constant learning rate with a decay at the end of training.

****
A Two-Stage Neural-Filter Pareto Front Extractor and the need for Benchmarking
Soumyajit Gupta,Gurpreet Singh,Matthew Lease

Pareto solutions are optimal trade-offs between multiple competing objectives over the feasible set satisfying imposed constraints. Fixed-point iterative strategies do not always converge and might only return one solution point per run. Co nsequently, multiple runs of a scalarization problem are required to retrieve a Pareto front, where all instances converge. Recently proposed Multi-Task Learning (MTL) solvers claim to achieve Pareto solutions combining Linear Scalarization and domain decomposition. We demonstrate key shortcomings of MTL solvers, that limit their usability for real-world applications. Issues include unjustified convexity assumptions on practical problems, incomplete and often wrong inferences on datasets that violate Pareto definition, and lack of proper benchmarking and verification. We propose a two stage Pareto framework: Hybrid Neural Pareto Front (HNPF) that is accurate and handles non-convex functions and constraints. The Stage-1 neural network efficiently extracts the \textit{weak} Pareto front, using Fritz-John Conditions (FJC) as the discriminator, with no assumptions of convexity on the objectives or constraints. An FJC guided diffusive manifold is used to bound the error between the true and the Stage-1 extracted \textit{weak} Pareto front. The Stage-2, low-cost Pareto filter then extracts the strong Pareto subset from this weak front. Numerical experiments demonstrates the accuracy and efficiency of our approach.

****

Sphere2Vec: Self-Supervised Location Representation Learning on Spherical Surfaces

Gengchen Mai,Yao Xuan,Wenyun Zuo,Yutong He,Stefano Ermon,Jiaming Song,Krzysztof Janowicz,Ni Lao

Location encoding is valuable for a multitude of tasks where both the absolute positions and local contexts (image, text, and other types of metadata) of spatial objects are needed for accurate predictions. However, most existing approaches do not leverage unlabeled data, which is crucial for use cases with limited labels. Furthermore, the availability of large-scale real-world GPS coordinate data demands representation and prediction at global scales. However, existing location encoding models assume that the input coordinates are in Euclidean space, which can lead to modeling errors due to distortions introduced when mapping coordinates from other manifolds (e.g., spherical surfaces) to Euclidean space. We introduceSphere2Vec, a location encoder, which can directly encode spherical coordinates while preserving spherical distances.Sphere2Vecis trained with a self-supervised learning framework which pre-trains deep location representations from unlabeled geo-tagged images with contrastive losses, and then fine-tunes to perform super-vised geographic object classification tasks.Sphere2Vecachieves the performances of state-of-the-art results on various image classification tasks ranging from species, Point of Interest (POI) facade, to remote sensing. The self-supervised pertaining significantly improves the performance ofSphere2Vecespecially when the labeled data is limited

****

The NTK Adversary: An Approach to Adversarial Attacks without any Model Access

Nikolaos Tsilivis,Julia Kempe

Adversarial attacks carefully perturb natural inputs, so that a machine learning algorithm produces erroneous decisions on them. Most successful attacks on neural networks exploit gradient information of the model (either directly or by estimating it through querying the model). Harnessing recent advances in Deep Learning theory, we propose a radically different attack that eliminates that need. In particular, in the regime where the Neural Tangent Kernel theory holds, we derive a simple, but powerful strategy for attacking models, which in contrast to prior work, does not require any access to the model under attack, or any trained replica of it for that matter. Instead, we leverage the explicit description afforded by the NTK to maximally perturb the output of the model, using solely information about the model structure and the training data. We experimentally verify the efficacy of our approach, first on models that lie close to the theoretical assumptions (large width, proper initialization, etc.) and, further, on more practical scenarios, with those assumptions relaxed. In addition, we show that our perturbations exhibit strong transferability between models.

****

Range-Net: A High Precision Neural SVD

Soumyajit Gupta,Gurpreet Singh,Clint N. Dawson

For Big Data applications, computing a rank-r Singular Value Decomposition (SVD) is restrictive due to the main memory requirements. Recently introduced streaming Randomized SVD schemes work under the restrictive assumption that the singular value spectrum of the data has an exponential decay. This is seldom true for any practical data. Further, the approximation errors in the singular vectors and values are high due to the randomized projection. We present Range-Net as a low memory alternative to rank-r SVD that satisfies the lower bound on tail-energy given by Eckart-Young-Mirsky (EYM) theorem at machine precision. Range-Net is a deterministic two-stage neural optimization approach with random initialization, where the memory requirement depends explicitly on the feature dimension and desired rank, independent of the sample dimension. The data samples are read in a streaming manner with the network minimization problem converging to the desired rank-r approximation. Range-Net is fully interpretable where all the network outputs and weights have a specific meaning. We provide theoretical guarantees that Range-Net extracted SVD factors satisfy EYM tail-energy lower bound with numerical experiments on real datasets at various scales that confirm these bounds. A comparison against the state-of-the-art streaming Randomized SVD shows that Range-Net is six orders of magnitude more accurate in terms of tail energy while correctly extracting the singular values and vectors.

****

Path Auxiliary Proposal for MCMC in Discrete Space

Haoran Sun,Hanjun Dai,Wei Xia,Arun Ramamurthy

Energy-based Model (EBM) offers a powerful approach for modeling discrete structure, but both inference and learning of EBM are hard as it involves sampling from discrete distributions. Recent work shows Markov Chain Monte Carlo (MCMC) with the informed proposal is a powerful tool for such sampling. However, an informed proposal only allows local updates as it requires evaluating all energy changes in the neighborhood.

In this work, we present a path auxiliary algorithm that uses a composition of local moves to efficiently explore large neighborhoods. We also give a fast version of our algorithm that only queries the evaluation of energy function twice for each proposal via linearization of the energy function. Empirically, we show that our path auxiliary algorithms considerably outperform other generic samplers on various discrete models for sampling, inference, and learning. Our method can also be used to train deep EBMs for high-dimensional discrete data.

****

Understanding Metric Learning on Unit Hypersphere and Generating Better Examples for Adversarial Training

Yihan Wu,Heng Huang

Recent works have shown that adversarial examples can improve the performance of representation learning tasks. In this paper, we boost the performance of deep metric learning (DML) models with adversarial examples generated by attacking two new objective functions: \textit{intra-class alignment} and \textit{hyperspherical uniformity}. These two new objectives come from our theoretical and empirical analysis of the tuple-based metric losses on the hyperspherical embedding space. Our analytical results reveal that a) the metric losses on positive sample pairs are related to intra-class alignment; b) the metric losses on negative sample pairs serve as uniformity regularization on hypersphere. Based on our new understanding on the DML models, we propose Adversarial Deep Metric Learning model with adversarial samples generated by Alignment or Uniformity objective (ADML+A or U). With the same network structure and training settings, ADML+A and ADML+U consistently outperform the state-of-the-art vanilla DML models and a baseline model, adversarial DML model with attacking triplet objective function, on four metric learning benchmarks.

****

An Agnostic Approach to Federated Learning with Class Imbalance

Zebang Shen,Juan Cervino,Hamed Hassani,Alejandro Ribeiro

Federated Learning (FL) has emerged as the tool of choice for training deep models over heterogeneous and decentralized datasets.

As a reflection of the experiences from different clients, severe class imbalanc e issues are observed in real-world FL problems.

Moreover, there exists a drastic mismatch between the imbalances from the local and global perspectives, i.e. a local majority class can be the minority of the population. Additionally, the privacy requirement of FL poses an extra challenge, as one should handle class imbalance without identifying the minority class. In this paper we propose a novel agnostic constrained learning formulation to tackle the class imbalance problem in FL, without requiring further information beyond the standard FL objective. A meta algorithm, CLIMB, is designed to solve the target optimization problem, with its convergence property analyzed under certain oracle assumptions. Through an extensive empirical study over various data heterogeneity and class imbalance configurations, we showcase that CLIMB considerably improves the performance in the minority class without compromising the overall accuracy of the classifier, which significantly outperforms previous arts.

In fact, we observe the greatest performance boost in the most difficult scenario where every client only holds data from one class. The code can be found here https://github.com/shenzebang/Federated-Learning-Pytorch.

****

NormFormer: Improved Transformer Pretraining with Extra Normalization

Sam Shleifer,Myle Ott

During pretraining, the Pre-LayerNorm transformer suffers from a gradient magnitude mismatch: gradients at early layers are much larger than at later layers, while the optimal weighting of residuals is larger at earlier than at later layers. These issues can be alleviated by the addition of two normalization and two new scaling operations inside each layer.

The extra operations incur negligible compute cost (+0.5\% parameter increase), but improve pretraining perplexity and downstream task performance for both causal and masked language models of multiple sizes.

Adding NormFormer on top of the GPT3-Medium architecture can reach the SOTA perplexity 22\% faster, or converge 0.33 perplexity better in the same compute budget. This results in significantly stronger zero shot performance.

For masked language modeling, NormFormer improves fine-tuned GLUE performance by 1.9\% on average.

****

A Fine-Tuning Approach to Belief State Modeling

Samuel Sokota,Hengyuan Hu,David J Wu,J Zico Kolter,Jakob Nicolaus Foerster,Noam Brown

We investigate the challenge of modeling the belief state of a partially observable Markov system, given sample-access to its dynamics model. This problem setting is often approached using parametric sequential generative modeling methods. However, these methods do not leverage any additional computation at inference time to increase their accuracy. Moreover, applying these methods to belief state modeling in certain multi-agent settings would require passing policies into the belief model---at the time of writing, there have been no successful demonstrations of this. Toward addressing these shortcomings, we propose an inference-time improvement framework for parametric sequential generative modeling methods called belief fine-tuning (BFT). BFT leverages approximate dynamic programming in the form of fine-tuning to determine the model parameters at each time step. It can improve the accuracy of the belief model at test time because it specializes the model to the space of local observations. Furthermore, because this specialization occurs after the action or policy has already been decided, BFT does not require the belief model to process it as input. As a result of the latter point, BFT enables, for the first time, approximate public belief state search in imperfect-information games where the number of possible information states is too large to track tabularly. We exhibit these findings on large-scale variants of the benchmark game Hanabi.

****

Quasi-Newton policy gradient algorithms

Haoya Li,Samarth Gupta,Hsiang-Fu Yu,Lexing Ying,Inderjit S Dhillon

Policy gradient algorithms have been widely applied to reinforcement learning (RL) problems in recent years. Regularization with various entropy functions is of ten used to encourage exploration and improve stability. In this paper, we propose a quasi-Newton method for the policy gradient algorithm with entropy regularization. In the case of Shannon entropy, the resulting algorithm reproduces the natural policy gradient (NPG) algorithm. For other entropy functions, this method results in brand new policy gradient algorithms. We provide a simple proof that all these algorithms enjoy the Newton-type quadratic convergence near the optimal policy. Using synthetic and industrial-scale examples, we demonstrate that the proposed quasi-Newton method typically converges in single-digit iterations, often orders of magnitude faster than other state-of-the-art algorithms.

****

Encoding Event-Based Gesture Data With a Hybrid SNN Guided Variational Auto-encoder

Kenneth Michael Stewart,Andreea Danielescu,Timothy Shea,Emre Neftci

Commercial mid-air gesture recognition systems have existed for at least a decade, but they have not become a widespread method of interacting with machines. These systems require rigid, dramatic gestures to be performed for accurate recognition that can be fatiguing and unnatural. To address this limitation, we propose a neuromorphic gesture analysis system which encodes event-based gesture data at high temporal resolution. Our novel approach consists of an event-based guided Variational Autoencoder (VAE) which encodes event-based data sensed by a Dynamic Vision Sensor (DVS) into a latent space representation suitable to compute the similarity of mid-air gesture data. We show that the Hybrid Guided-VAE achieves 87% classification accuracy on the DVSGesture dataset and it can encode the sparse, noisy inputs into an interpretable latent space representation, visualized through T-SNE plots. We also implement the encoder component of the model on neuromorphic hardware and discuss the potential for our algorithm to enable real-time, self-supervised learning of natural mid-air gestures.

****

Differentially Private Fine-tuning of Language Models

Da Yu,Saurabh Naik,Arturs Backurs,Sivakanth Gopi,Huseyin A Inan,Gautam Kamath,Janardhan Kulkarni,Yin Tat Lee,Andre Manoel,Lukas Wutschitz,Sergey Yekhanin,Huishuai Zhang

We give simpler, sparser, and faster algorithms for differentially private fine-tuning of large-scale pre-trained language models, which achieve the state-of-the-art privacy versus utility tradeoffs on many standard NLP tasks. We propose a meta-framework for this problem, inspired by the recent success of highly parameter-efficient methods for fine-tuning. Our experiments show that differentially private adaptations of these approaches outperform previous private algorithms in three important dimensions: utility, privacy, and the computational and memory cost of private training. On many commonly studied datasets, the utility of private models approaches that of non-private models. For example, on the MNLI dataset we achieve an accuracy of 87.8% using RoBERTa-Large and 83.5% using RoBERTa-Base with a privacy budget of $\epsilon = 6.7$. In comparison, absent privacy constraints, RoBERTa-Large achieves an accuracy of 90.2%. Our findings are similar for natural language generation when privately fine-tuning GPT-2. Our experiments also show that larger models are better suited for private fine-tuning: while they are well known to achieve superior accuracy non-privately, we find that they also better maintain their accuracy when privacy is introduced.

****

Adversarial Weight Perturbation Improves Generalization in Graph Neural Networks

Yihan Wu,Aleksandar Bojchevski,Heng Huang

There is growing theoretical and empirical evidence that flatter local minima tend to improve generalization. An efficient and effective technique for finding such minima is Adversarial Weight Perturbation (AWP). The main idea is to minimize the loss w.r.t. a bounded worst-case perturbation of the model parameters by (approximately) solving an associated min-max problem. Intuitively, we favor local minima with a small loss in a neighborhood around them. The benefits of AWP, and more generally the connections between flatness and generalization, have been extensively studied for i.i.d. data such as images. In this paper we initiate the first study of this phenomenon for graph data. Along the way, we identify a v anishing-gradient issue with all existing formulations of AWP and we propose Weighted Truncated AWP (WT-AWP) to alleviate this issue. We show that regularizing graph neural networks with WT-AWP consistently improves both natural and robust generalization across many different graph learning tasks and models.

****

Equalized Robustness: Towards Sustainable Fairness Under Distributional Shifts

Haotao Wang,Junyuan Hong,Jiayu Zhou,Zhangyang Wang

Increasing concerns have been raised on deep learning fairness in recent years. Existing fairness metrics and algorithms mainly focus on the discrimination of model performance across different groups on in-distribution data. It remains unclear whether the fairness achieved on in-distribution data can be generalized to data with unseen distribution shifts, which are commonly encountered in real-world applications. In this paper, we first propose a new fairness goal, termed Equalized Robustness (ER), to impose fair model robustness against unseen distribution shifts across majority and minority groups. ER measures robustness disparity by the maximum mean discrepancy (MMD) distance between the loss curvature distributions of two groups of data. We show that previous fairness learning algorithms designed for in-distribution fairness fail to meet the new robust fairness goal. We further propose a novel fairness learning algorithm, termed Curvature Matching (CUMA), to simultaneously achieve both traditional in-distribution fairness and our new robust fairness. CUMA efficiently debiases the model robustness by minimizing the MMD distance between loss curvature distributions of two groups. Experiments on three popular datasets show CUMA achieves superior fairness in robustness against distribution shifts, without more sacrifice on either overall accuracies or the in-distribution fairness.

****

P-Adapters: Robustly Extracting Factual Information from Language Models with Diverse Prompts

Benjamin Newman,Prafulla Kumar Choubey,Nazneen Rajani

Recent work (e.g. LAMA (Petroni et al., 2019)) has found that the quality of the factual information extracted from Large Language Models (LLMs) depends on the prompts used to query them. This inconsistency is problematic because different users will query LLMs for the same information using different wording, but should receive the same, accurate responses regardless. In this work we aim to address this shortcoming by introducing P-Adapters: lightweight models that sit between the embedding layer and first attention layer of LLMs. They take LLM embeddings as input and output continuous prompts that are used to query the LLM. Additionally, we investigate Mixture of Experts (MoE) models that learn a set of continuous prompts (the "experts") and select one to query the LLM. These require a separate classifier trained on human-annotated data to map natural language prompts to the continuous ones. P-Adapters perform comparably to the more complex MoE models in extracting factual information from BERT and RoBERTa while eliminating the need for additional annotations. P-Adapters show between 12-26% absolute improvement in precision and 36-50% absolute improvement in consistency over a baseline of just using natural language queries alone. Finally, we investigate what makes P-Adapters successful and conclude that a significant factor is access to the LLM's embeddings of the original natural language prompt, particularly the subject of the entity pair being queried.

****

Iterated Reasoning with Mutual Information in Cooperative and Byzantine Decentralized Teaming

Sachin G Konan,Esmaeil Seraj,Matthew Gombolay

Information sharing is key in building team cognition and enables coordination and cooperation. High-performing human teams also benefit from acting strategically with hierarchical levels of iterated communication and rationalizability, meaning a human agent can reason about the actions of their teammates in their decision-making. Yet, the majority of prior work in Multi-Agent Reinforcement Learning (MARL) does not support iterated rationalizability and only encourage inter-agent communication, resulting in a suboptimal equilibrium cooperation strategy. In this work, we show that reformulating an agent's policy to be conditional on the policies of its neighboring teammates inherently maximizes Mutual Information (MI) lower-bound when optimizing under Policy Gradient (PG). Building on the idea of decision-making under bounded rationality and cognitive hierarchy theory, we show that our modified PG approach not only maximizes local agent rewards but also implicitly reasons about MI between agents without the need for any explicit ad-hoc regularization terms. Our approach, InfoPG, outperforms baselines in learning emergent collaborative behaviors and sets the state-of-the-art in decentralized cooperative MARL tasks. Our experiments validate the utility of InfoPG by achieving higher sample efficiency and significantly larger cumulative reward in several complex cooperative multi-agent domains.

****

Towards Generalizable Personalized Federated Learning with Adaptive Local Adaptation
Sijia Chen,Baochun Li

Personalized federated learning aims to find a shared global model that can be adapted to meet personal needs on each individual device. Starting from such a shared initial model, devices should be able to easily adapt to their local dataset to obtain personalized models. However, we find that existing works cannot generalize well on non-iid scenarios with different heterogeneity degrees of the underlying data distribution among devices. Thus, it is challenging for these methods to train a suitable global model to effectively induce high-quality personalized models without changing learning objectives. In this paper, we point out that this issue can be addressed by balancing information flow from the initial model and training dataset to the local adaptation. We then prove a theorem referred to as the {\em adaptive trade-off theorem}, showing adaptive local adaptation is equivalent to optimizing such information flow based on the information theory. With these theoretical insights, we propose a new framework called {\em adaptive federated meta-learning} (AFML), designed to achieve generalizable personalized federated learning that maintains solid performance under non-IID data scenarios with different degrees of diversity among devices. We test AFML in an extensive set of these non-IID data scenarios, with both CIFAR-100 and Shakespeare datasets. Experimental results demonstrate that AFML can maintain the highest personalized accuracy compared to alternative leading frameworks, yet with a minimal number of communication rounds and local updates needed.

****

Possibility Before Utility: Learning And Using Hierarchical Affordances
Robby Costales,Shariq Iqbal,Fei Sha

Reinforcement learning algorithms struggle on tasks with complex hierarchical dependency structures. Humans and other intelligent agents do not waste time assessing the utility of every high-level action in existence, but instead only consider ones they deem possible in the first place. By focusing only on what is feasible, or "afforded'', at the present moment, an agent can spend more time both evaluating the utility of and acting on what matters. To this end, we present Hierarchical Affordance Learning (HAL), a method that learns a model of hierarchical affordances in order to prune impossible subtasks for more effective learning. Existing works in hierarchical reinforcement learning provide agents with structural representations of subtasks but are not affordance-aware, and by grounding our definition of hierarchical affordances in the present state, our approach is more flexible than the multitude of approaches that ground their subtask dependencies in a symbolic history. While these logic-based methods often require complete knowledge of the subtask hierarchy, our approach is able to utilize incomplete and varying symbolic specifications. Furthermore, we demonstrate that relative to non-affordance-aware methods, HAL agents are better able to efficiently learn complex tasks, navigate environment stochasticity, and acquire diverse skills in the absence of extrinsic supervision---all of which are hallmarks of human learning.

****

Layer-wise Adaptive Model Aggregation for Scalable Federated Learning
Sunwoo Lee,Tuo Zhang,Chaoyang He,Salman Avestimehr

In Federated Learning, a common approach for aggregating local models across cli ents is periodic averaging of the full model parameters. It is, however, known that different layers of neural networks can have a different degree of model discrepancy across the clients. The conventional full aggregation scheme does not consider such a difference and synchronizes the whole model parameters at once, resulting in inefficient network bandwidth consumption. Aggregating the parameters that are similar across the clients does not make meaningful training progress while increasing the communication cost. We propose FedLAMA, a layer-wise model aggregation scheme for scalable Federated Learning. FedLAMA adaptively adjusts the aggregation interval in a layer-wise manner, jointly considering the model discrepancy and the communication cost. The layer-wise aggregation method enables to finely control the aggregation interval to relax the aggregation frequency without a significant impact on the model accuracy. Our empirical study shows that FedLAMA reduces the communication cost by up to 60% for IID data and 70% for non-IID data while achieving a comparable accuracy to FedAvg.

****

DFSSATTEN: Dynamic Fine-grained Structured Sparse Attention Mechanism
Zhaodong Chen,Liu Liu,Yuying Quan,Zheng Qu,Yufei Ding,Yuan Xie

Transformers are becoming mainstream solutions for various tasks like NLP and Computer vision. Despite their success, the quadratic complexity of their attention mechanism hinders them from applying to latency sensitive tasks. Tremendous efforts have been made to alleviate this problem, and many of them successfully reduce the asymptotic complexity to linear. Nevertheless, few of them achieve practical speedup over the original full attention, especially under the moderate sequence length. In this paper, we present DFSSATTEN, an attention mechanism that dynamically prunes the full attention weight matrix to the 50% fine-grained structured sparse pattern used by the sparse tensor core on NVIDIA A100 GPU. We provide both theoretical and empirical evidences that demonstrate DFSSAT- TEN is a good approximation of the full attention mechanism and can achieve speedups in wall-clock time under arbitrary sequence length. We evaluate our method on tasks from various domains under different sequence lengths from 256 to 4096. DFSSATTEN achieves 1.27 ~ 1.89× speedups over the full-attention mechanism with no accuracy loss.

****

Step-unrolled Denoising Autoencoders for Text Generation
Nikolay Savinov,Junyoung Chung,Mikolaj Binkowski,Erich Elsen,Aaron van den Oord

In this paper we propose a new generative model of text, Step-unrolled Denoising Autoencoder (SUNDAE), that does not rely on autoregressive models. Similarly to denoising diffusion techniques, SUNDAE is repeatedly applied on a sequence of tokens, starting from random inputs and improving them each time until convergence. We present a simple new improvement operator that converges in fewer iterations than diffusion methods, while qualitatively producing better samples on natural language datasets. SUNDAE achieves state-of-the-art results (among non-autoregressive methods) on the WMT'14 English-to-German translation task and good qualitative results on unconditional language modeling on the Colossal Cleaned Common Crawl dataset and a dataset of Python code from GitHub. The non-autoregressive nature of SUNDAE opens up possibilities beyond left-to-right prompted generation, by filling in arbitrary blank patterns in a template.

****

Loss Function Learning for Domain Generalization by Implicit Gradient
Boyan Gao,Henry Gouk,Yongxin Yang,Timothy Hospedales

Generalising robustly to distribution shift is a major challenge that is pervasive across most real-world applications of machine learning. A recent study highlighted that many advanced algorithms proposed to tackle such domain generalisation (DG) fail to outperform a properly tuned empirical risk minimisation (ERM) baseline. We take a different approach, and explore the impact of the ERM loss function on out-of-domain generalisation. In particular, we introduce a novel meta-learning approach to loss function search based on implicit gradient. This enables us to discover a general purpose parametric loss function that provides a drop-in replacement for cross-entropy. Our loss can be used in standard training pipelines to efficiently train robust models using any neural architecture on new datasets. The results show that it clearly surpasses cross-entropy, enables simple ERM to outperform significantly more complicated prior DG methods, and provides state-of-the-art performance across a variety of DG benchmarks. Furthermore, unlike most existing DG approaches, our setup applies to the most practical setting of single-source domain generalisation, on which we show significant improvement.

****

Hindsight Foresight Relabeling for Meta-Reinforcement Learning

Michael Wan,Jian Peng,Tanmay Gangwani

Meta-reinforcement learning (meta-RL) algorithms allow for agents to learn new behaviors from small amounts of experience, mitigating the sample inefficiency problem in RL. However, while meta-RL agents can adapt quickly to new tasks at test time after experiencing only a few trajectories, the meta-training process is still sample-inefficient. Prior works have found that in the multi-task RL setting, relabeling past transitions and thus sharing experience among tasks can improve sample efficiency and asymptotic performance. We apply this idea to the meta-RL setting and devise a new relabeling method called Hindsight Foresight Relabeling (HFR). We construct a relabeling distribution using the combination of "hindsight", which is used to relabel trajectories using reward functions from the training task distribution, and "foresight", which takes the relabeled trajectories and computes the utility of each trajectory for each task. HFR is easy to implement and readily compatible with existing meta-RL algorithms. We find that HFR improves performance when compared to other relabeling methods on a variety of meta-RL tasks.

****

Multi-Domain Self-Supervised Learning

Neha Mukund Kalibhat,Yogesh Balaji,C. Bayan Bruss,Soheil Feizi

Contrastive self-supervised learning has recently gained significant attention owing to its ability to learn improved feature representations without the use of label information. Current contrastive learning approaches, however, are only effective when trained on a particular dataset, limiting their utility in diverse multi-domain settings. In fact, training these methods on a combination of several domains often degrades the quality of learned representations compared to the models trained on a single domain. In this paper, we propose a Multi-Domain Self-Supervised Learning (MDSSL) approach that can effectively perform representation learning on multiple, diverse datasets. In MDSSL, we propose a three-level hierarchical loss for measuring the agreement between augmented views of a given sample, agreement between samples within a dataset and agreement between samples across datasets. We show that MDSSL when trained on a mixture of CIFAR-10, STL-10, SVHN and CIFAR-100 produces powerful representations, achieving up to a 25% increase in top-1 accuracy on a linear classifier compared to single-domain self-supervised encoders. Moreover, MDSSL encoders can generalize more effectively to unseen datasets compared to both single-domain and multi-domain baselines. MDSSL is also highly efficient in terms of the resource usage as it stores and trains a single model for multiple datasets leading up to 17% reduction in training time. Finally, for multi-domain datasets where domain labels are unknown, we propose a modified approach that alternates between clustering and MDSSL. Thus, for diverse multi-domain datasets (even without domain labels), MDSSL provides an efficient and generalizable self-supervised encoder without sacrificing the quality of representations in individual domains.

****

An Application of Pseudo-log-likelihoods to Natural Language Scoring

Darren Abramson,Ali Emami

Language models built using semi-supervised machine learning on large corpora of natural language have very quickly enveloped the fields of natural language generation and understanding. In this paper we apply a zero-shot approach in- dependently developed by several researchers now gaining recognition as a significant alternative to fine-tuning for evaluation on common sense tasks. A language model with relatively few parameters and training steps (albert-xxlarge-v2) compared to a more recent language model (T5) can outperform it on a recent large data set (TimeDial), while displaying robustness in its performance across a similar class of language tasks. Surprisingly, this result is achieved by using a hyperparameter-free zero-shot method with the smaller model, compared to fine-tuning to the larger model. We argue that robustness of the smaller model ought to be understood in terms of compositionality, in a sense that we draw from re- cent literature on a class of similar models. We identify a practical cost for our method and model: high GPU-time for natural language evaluation. The zero-shot measurement technique that produces remarkable stability, both for ALBERT and other BERT variants, is an application of pseudo-log-likelihoods to masked language models for the relative measurement of probability for substitution alter- natives in forced choice language tasks such as the Winograd Schema Challenge, Winogrande, CommonsenseQA, and others. One contribution of this paper is to bring together a number of similar, but independent strands of research. We produce some absolute state-of-the-art (SOTA) results for common sense reasoning in binary choice tasks, performing better than any published result in the literature, including fine-tuned efforts. In others our results are SOTA relative to published methods similar to our own – in some cases by wide margins, but below SOTA absolute for fine-tuned alternatives. In addition, we show a remarkable consistency of the model’s performance under adversarial settings, which we argue is best explained by the model’s compositionality of representations.

****

Human-Level Control without Server-Grade Hardware

Brett Daley,Christopher Amato

Deep Q-Network (DQN) marked a major milestone for reinforcement learning, demonstrating for the first time that human-level control policies could be learned directly from raw visual inputs via reward maximization. Even years after its introduction, DQN remains highly relevant to the research community since many of its innovations have been adopted by successor methods. Nevertheless, despite significant hardware advances in the interim, DQN's original Atari 2600 experiments remain extremely costly to replicate in full. This poses an immense barrier to researchers who cannot afford state-of-the-art hardware or lack access to large-scale cloud computing resources. To facilitate improved access to deep reinforcement learning research, we introduce a DQN implementation that leverages a novel concurrent and synchronized execution framework designed to maximally utilize a heterogeneous CPU-GPU desktop system. With just one NVIDIA GeForce GTX 1080 GPU, our implementation reduces the training time of a 200-million-frame Atari experiment from 25 hours to just 9 hours. The ideas introduced in our paper should be generalizable to a large number of off-policy deep reinforcement learning methods.

****

Fragment-Based Sequential Translation for Molecular Optimization

Benson Chen,Xiang Fu,Tommi S. Jaakkola,Regina Barzilay

Searching for novel molecular compounds with desired properties is an important problem in drug discovery. Many existing frameworks generate molecules one atom at a time. We instead propose a flexible editing paradigm that generates molecules using learned molecular fragments---meaningful substructures of molecules. To do so, we train a variational autoencoder (VAE) to encode molecular fragments in a coherent latent space, which we then utilize as a vocabulary for editing molecules to explore the complex chemical property space. Equipped with the learned fragment vocabulary, we propose Fragment-based Sequential Translation (FaST), which learns a reinforcement learning (RL) policy to iteratively translate model-discovered molecules into increasingly novel molecules while satisfying desired properties. Empirical evaluation shows that FaST significantly improves over state-of-the-art methods on benchmark single/multi-objective molecular optimization tasks.

****

Learning Temporally-Consistent Representations for Data-Efficient Reinforcement Learning

Trevor McInroe,Lukas Schäfer,Stefano V Albrecht

Deep reinforcement learning (RL) agents that exist in high-dimensional state spa ces, such as those composed of images, have interconnected learning burdens. Agents must learn an action-selection policy that completes their given task, which requires them to learn a representation of the state space that discerns between useful and useless information. The reward function is the only supervised feedback that RL agents receive, which causes a representation learning bottleneck that can manifest in poor sample efficiency. We present k-Step Latent (KSL), a new representation learning method that enforces temporal consistency of representations via a self-supervised auxiliary task wherein agents learn to recurrently predict action-conditioned representations of the state space. The state encoder learned by KSL produces low-dimensional representations that make optimization of the RL task more sample efficient. Altogether, KSL produces state-of-the-art results in both data efficiency and asymptotic performance in the popular PlaNet benchmark suite. Our analyses show that KSL produces encoders that generalize better to new tasks unseen during training, and its representations are more strongly tied to reward, are more invariant to perturbations in the state space, and move more smoothly through the temporal axis of the RL problem than other methods such as DrQ, RAD, CURL, and SAC-AE.

****

Training Data Size Induced Double Descent For Denoising Neural Networks and the Role of Training Noise Level

Rishi Sonthalia,Raj Rao Nadakuditi

When training a denoising neural network, we show that more data isn’t more beneficial. In fact the generalization error versus number of of training data points is a double descent curve.

Training a network to denoise noisy inputs is the most widely used technique for pre-training deep neural networks. Hence one important question is the effect of scaling the number of training data points. We formalize the question of how many data points should be used by looking at the generalization error for denoising noisy test data. Prior work on computing the generalization error focus on adding noise to target outputs. However, adding noise to the input is more in line with current pre-training practices. In the linear (in the inputs) regime, we provide an asymptotically exact formula for the generalization error for rank 1 data and an approximation for the generalization error for rank r data. We show using our formulas, that the generalization error versus number of data points follows a double descent curve. From this, we derive a formula for the amount of noise that needs to be added to the training data to minimize the denoising error and see that this follows a double descent curve as well.

****

Interpretable Unsupervised Diversity Denoising and Artefact Removal

Mangal Prakash,Mauricio Delbracio,Peyman Milanfar,Florian Jug

Image denoising and artefact removal are complex inverse problems admitting multiple valid solutions. Unsupervised diversity restoration, that is, obtaining a diverse set of possible restorations given a corrupted image, is important for ambiguity removal in many applications such as microscopy where paired data for supervised training are often unobtainable. In real world applications, imaging noise and artefacts are typically hard to model, leading to unsatisfactory performance of existing unsupervised approaches. This work presents an interpretable approach for unsupervised and diverse image restoration. To this end, we introduce a capable architecture called Hierarchical DivNoising (HDN) based on hierarchical Variational Autoencoder. We show that HDN learns an interpretable multi-scale representation of artefacts and we leverage this interpretability to remove imaging artefacts commonly occurring in microscopy data. Our method achieves state-of-the-art results on twelve benchmark image denoising datasets while providing access to a whole distribution of sensibly restored solutions.

Additionally, we demonstrate on three real microscopy datasets that HDN removes artefacts without supervision, being the first method capable of doing so while generating multiple plausible restorations all consistent with the given corrupted image.

****

LoRA: Low-Rank Adaptation of Large Language Models

Edward J Hu,yelong shen,Phillip Wallis,Zeyuan Allen-Zhu,Yuanzhi Li,Shean Wang,Lu Wang,Weizhu Chen

An important paradigm of natural language processing consists of large-scale pre-training on general domain data and adaptation to particular tasks or domains. As we pre-train larger models, full fine-tuning, which retrains all model parameters, becomes less feasible.

Using GPT-3 175B as an example -- deploying independent instances of fine-tuned models, each with 175B parameters, is prohibitively expensive. We propose Low-Rank Adaptation, or LoRA, which freezes the pre-trained model weights and injects trainable rank decomposition matrices into each layer of the Transformer architecture, greatly reducing the number of trainable parameters for downstream tasks. Compared to GPT-3 175B fine-tuned with Adam, LoRA can reduce the number of trainable parameters by a factor of 10,000 and the GPU memory requirement by a factor of 3. LoRA performs on-par or better than fine-tuning in model quality on RoBERTa, DeBERTa, GPT-2, and GPT-3, despite having fewer trainable parameters, a higher training throughput, and, unlike adapters, no additional inference latency. We also provide an empirical investigation into rank-deficiency in language model adaptation, which sheds light on the efficacy of LoRA. We release a package that facilitates the integration of LoRA with PyTorch models and provide our implementations and model checkpoints for RoBERTa, DeBERTa, and GPT-2 at https://github.com/microsoft/LoRA.

****

Which Shortcut Cues Will DNNs Choose? A Study from the Parameter-Space Perspective

Luca Scimeca,Seong Joon Oh,Sanghyuk Chun,Michael Poli,Sangdoo Yun

Deep neural networks (DNNs) often rely on easy–to–learn discriminatory features, or cues, that are not necessarily essential to the problem at hand. For example, ducks in an image may be recognized based on their typical background scenery, such as lakes or streams. This phenomenon, also known as shortcut learning, is emerging as a key limitation of the current generation of machine learning models. In this work, we introduce a set of experiments to deepen our understanding of shortcut learning and its implications. We design a training setup with several shortcut cues, named WCST-ML, where each cue is equally conducive to the visual recognition problem at hand. Even under equal opportunities, we observe that (1) certain cues are preferred to others, (2) solutions biased to the easy–to–learn cues tend to converge to relatively flat minima on the loss surface, and (3) the solutions focusing on those preferred cues are far more abundant in the parameter space. We explain the abundance of certain cues via their Kolmogorov (descriptional) complexity: solutions corresponding to Kolmogorov-simple cues are abundant in the parameter space and are thus preferred by DNNs. Our studies are based on the synthetic dataset DSprites and the face dataset UTKFace. In our WCST-ML, we observe that the inborn bias of models leans toward simple cues, such as color and ethnicity. Our findings emphasize the importance of active human intervention to remove the inborn model biases that may cause negative societal impacts.

****

Unifying Distribution Alignment as a Loss for Imbalanced Semi-supervised Learning

Justin Lazarow,Kihyuk Sohn,Chun-Liang Li,Zizhao Zhang,Chen-Yu Lee,Tomas Pfister

While remarkable progress in imbalanced supervised learning has been made recently, less attention has been given to the setting of imbalanced semi-supervised learning (SSL) where not only is a few labeled data provided, but the underlying data distribution can be severely imbalanced. Recent works require both complicated sampling-based strategies of pseudo-labeled data and distribution alignment of the pseudo-label distribution to accommodate this imbalance. We present a novel approach that relies only on a form of a distribution alignment but no sampling strategy where rather than aligning the pseudo-labels during inference, we move the distribution alignment component into the respective cross entropy loss computations for both the supervised and unsupervised losses. This alignment compensates for both imbalance in the data as well as the eventual distributional sh ift present during evaluation. Altogether, this provides a single, unified strategy that offers both significantly reduced training requirements and improved performance across both low and richly labeled regimes and over varying degrees of imbalance. In experiments, we validate the efficacy of our method on SSL variants of CIFAR10-LT, CIFAR100-LT, and ImageNet-127. On ImageNet-127, our method shows 1.6% accuracy improvement over the previous best method with 80% training time reduction.

****

Efficient Packing: Towards 2x NLP Speed-Up without Loss of Accuracy for BERT

Matej Kosec,Sheng Fu,Mario Michael Krell

We find that at sequence length 512 padding tokens represent in excess of 50% of the Wikipedia dataset used for pretraining BERT (Bidirectional Encoder Representations from Transformers). Therefore by removing all padding we achieve a 2x speed-up in terms of sequences/sec. To exploit this characteristic of the dataset, we develop and contrast two deterministic packing algorithms. Both algorithms rely on the assumption that sequences are interchangeable and therefore packing can be performed on the histogram of sequence lengths, rather than per sample. This transformation of the problem leads to algorithms which are fast and have linear complexity in dataset size. The shortest-pack-first histogram-packing (SPFHP) algorithm determines the packing order for the Wikipedia dataset of over 16M sequences in 0.02 seconds. The non-negative least-squares histogram-packing (NNLSHP) algorithm converges in 28.4 seconds but produces solutions which are more depth efficient, managing to get near optimal packing by combining a maximum of 3 sequences in one sample. Using the dataset with multiple sequences per sample requires adjusting the model and the hyperparameters. We demonstrate that these changes are straightforward to implement and have relatively little impact on the achievable performance gain on modern hardware. Finally, we pretrain BERT-Large using the packed dataset, demonstrating no loss of convergence and the desired 2x speed-up.

****

Sequoia: A Software Framework to Unify Continual Learning Research

Fabrice Normandin,Oleksiy Ostapenko,Pau Rodriguez,Florian Golemo,Ryan Lindeborg,Matthew D Riemer,Lucas Cecchi,Timothee LESORT,Khimya Khetarpal,David Vazquez,Laurent Charlin,Irina Rish,Massimo Caccia

The field of Continual Learning (CL) seeks to develop algorithms that accumulate knowledge and skills over time through interaction with non-stationary environments. In practice, a plethora of evaluation procedures (settings) and algorithmic solutions (methods) exist, each with their own potentially disjoint set of assumptions. This variety makes measuring progress in CL difficult. We propose a taxonomy of settings, where each setting is described as a set of assumptions. A tree-shaped hierarchy emerges from this view, where more general settings become the parents of those with more restrictive assumptions. This makes it possible to use inheritance to share and reuse research, as developing a method for a given setting also makes it directly applicable onto any of its children. We instantiate this idea as a publicly available software framework called Sequoia, which features a wide variety of settings from both the Continual Supervised Learning (CSL) and Continual Reinforcement Learning (CRL) domains. Sequoia also includes a growing suite of methods which are easy to extend and customize, in addition to more specialized methods from external libraries. We hope that this new paradigm and its first implementation can help unify and accelerate research in CL. You can help us grow the tree by visiting (this GitHub URL).

****

Avoiding Overfitting to the Importance Weights in Offline Policy Optimization

Yao Liu,Emma Brunskill

Offline policy optimization has a critical impact on many real-world decision-making problems, as online learning is costly and concerning in many applications. Importance sampling and its variants are a widely used type of estimator in offline policy evaluation, which can be helpful to remove assumptions on the chosen function approximations used to represent value functions and process models. In this paper, we identify an important overfitting phenomenon in optimizing the importance weighted return, and propose an algorithm to avoid this overfitting. We provide a theoretical justification of the proposed algorithm through a better per-state-neighborhood normalization condition and show the limitation of previous attempts to this approach through an illustrative example. We further test our proposed method in a healthcare-inspired simulator and a logged dataset collected from real hospitals. These experiments show the proposed method with less overfitting and better test performance compared with state-of-the-art batch reinforcement learning algorithms.

****

Design and Evaluation for Robust Continual Learning

Yeu-Shin Fu,Josh Milthorpe

Continual learning is the ability to learn from new experiences without forgetting previous experiences. Different continual learning methods are each motivated by their own interpretation of the continual learning scenario, resulting in a wide variety of experiment protocols, which hinders understanding and comparison of results. Existing works emphasize differences in accuracy without considering the effects of experimental settings. However, understanding the effects of experimental assumptions is the most crucial part of any evaluation, as the experimental protocol may supply implicit information. We propose six rules as a guideline for experimental design and execution to conduct robust continual learning evaluation for better understanding of the methods. Using these rules, we demonstrate the importance of experimental choices regarding the sequence of incoming data and the sequence of the task oracle. Even when task oracle-based methods are desired, the rules can guide experimental design to support better evaluation and understanding of the continual learning methods. Consistent application of these rules in evaluating continual learning methods makes explicit the effect and validity of many assumptions, thereby avoiding misleading conclusions.

****

Cross-Architecture Distillation Using Bidirectional CMOW Embeddings

Lukas Paul Achatius Galke,Isabelle Cuber,Christoph Meyer,Henrik Ferdinand Nölscher,Angelina Sonderecker,Ansgar Scherp

Large pretrained language models (PreLMs) are revolutionizing natural language processing across all benchmarks. However, their sheer size is prohibitive for small laboratories or deployment on mobile devices. Approaches like pruning and distillation reduce the model size but typically retain the same model architecture. In contrast, we explore distilling PreLMs into a different, more efficient architecture CMOW, which embeds each word as a matrix and uses matrix multiplication to encode sequences. We extend the CMOW architecture and its CMOW/CBOW-Hybrid variant with a bidirectional component, per-token representations for distillation during pretraining, and a two-sequence encoding scheme that facilitates downstream tasks on sentence pairs such as natural language inferencing. Our results show that the embedding-based models yield scores comparable to DistilBERT on QQP and RTE, while using only half of its parameters and providing three times faster inference speed. We match or exceed the scores of ELMo, and only fall behind more expensive models on linguistic acceptability. Still, our distilled bidirectional CMOW/CBOW-Hybrid model more than doubles the scores on linguistic acceptability compared to previous cross-architecture distillation approaches. Furthermore, our experiments confirm the positive effects of bidirection and the two-sequence encoding scheme.

****

SAFER: Data-Efficient and Safe Reinforcement Learning Through Skill Acquisition

Dylan Z Slack,Yinlam Chow,Bo Dai,Nevan Wichers

Though many reinforcement learning (RL) problems involve learning policies in settings that are difficult to specify safety constraints and sparse rewards, current methods struggle to rapidly and safely acquire successful policies. Behavioral priors, which extract useful policy primitives for learning from offline datasets, have recently shown considerable promise at accelerating RL in more complex problems. However, we discover that current behavioral priors may not be well-equipped for safe policy learning, and in some settings, may promote unsafe behavior, due to their tendency to ignore data from undesirable behaviors. To overcome these issues, we propose SAFEty skill pRiors (SAFER), a behavioral prior learning algorithm that accelerates policy learning on complex control tasks, under safety constraints. Through principled contrastive training on safe and unsafe data, SAFER learns to extract a safety variable from offline data that encodes safety requirements, as well as the safe primitive skills over abstract actions in different scenarios. In the inference stage, SAFER composes a safe and successful policy from the safety skills according to the inferred safety variable and abstract action. We demonstrate its effectiveness on several complex safety-critical robotic grasping tasks inspired by the game Operation, in which SAFER not only out-performs baseline methods in learning successful policies but also enforces safety more effectively.

****

Monotonicity as a requirement and as a regularizer: efficient methods and applications

Joao Monteiro,Mohamed Osama Ahmed,Hossein Hajimirsadeghi,Greg Mori

We study the setting where risk minimization is performed over general classes of models and consider two cases where monotonicity is treated as either a requirement to be satisfied everywhere or a useful property. We specifically consider cases where point-wise gradient penalties are used alongside the empirical risk during training. In our first contribution, we show that different choices of penalties define the regions of the input space where the property is observed. As such, previous methods result in models that are monotonic only in a small volume of the input space. We thus propose an approach that uses mixtures of training instances and random points to populate the space and enforce the penalty in a much larger region. As a second contribution, we introduce the notion of monotonicity as a regularization bias for convolutional models. In this case, we consider applications, such as image classification and generative modeling, where monotonicity is not a hard constraint but can help improve some aspects of the model. Namely, we show that using group monotonicity can be beneficial in several applications such as: (1) defining strategies to detect anomalous data, (2) allowing for controllable data generation, and (3) generating explanations for predictions. Our proposed approaches do not introduce relevant computational overhead while leading to efficient procedures that provide extra benefits over baseline models.

****

Efficient Computation of Deep Nonlinear Infinite-Width Neural Networks that Learn Features

Greg Yang,Michael Santacroce,Edward J Hu

While a popular limit of infinite-width neural networks, the Neural Tangent Kernel (NTK) often exhibits performance gaps from finite-width neural networks on standard datasets, due to lack of feature learning. Although the feature learning *maximal update limit*, or *μ-limit* (Yang and Hu, 2020) of wide networks has closed the gap for 1-hidden-layer linear models, no one has been able to demonstrate this for deep nonlinear multi-layer perceptrons (MLP) because of μ-limit’s computational difficulty in this setting. Here, we solve this problem by proposing a novel feature learning limit, the *π-limit*, that bypasses the computational issues. The π-limit, in short, is the limit of a form of projected gradient descent, and the π-limit of an MLP is roughly another MLP where gradients are appended to weights during training. We prove its almost sure convergence with width using the Tensor Programs technique. We evaluate it on CIFAR10 and Omniglot against NTK as well as finite networks, finding the π-limit outperform finite-width models trained normally (without projection) in both settings, closing the perfo rmance gap between finite- and infinite-width neural networks previously left by NTK. Code for this work is available at github.com/santacml/pilim.

****

TRAIL: Near-Optimal Imitation Learning with Suboptimal Data
Mengjiao Yang,Sergey Levine,Ofir Nachum

In imitation learning, one aims to learn task-solving policies using access to near-optimal expert trajectories collected from the task environment. However, high-quality trajectories -- e.g., from human experts -- can be expensive to obtain in practical settings. On the contrary, it is often much easier to obtain large amounts of suboptimal trajectories which can nevertheless provide insight into the structure of the environment, showing what \emph{could} be done in the environment even if not what \emph{should} be done. Is it possible to formalize these conceptual benefits and devise algorithms to use offline datasets to yield \emph{provable} improvements to the sample-efficiency of imitation learning? In this work, we answer this question affirmatively and present training objectives which use an offline dataset to learn an approximate \emph{factored} dynamics model whose structure enables the extraction of a \emph{latent action space}. Our theoretical analysis shows that the learned latent action space can boost the sample-efficiency of downstream imitation learning, effectively reducing the need for large near-optimal expert datasets through the use of auxiliary non-expert data. We evaluate the practicality of our objective through experiments on a set of navigation and locomotion tasks. Our results verify the benefits suggested by our theory and show that our algorithms is able to recover near-optimal policies with fewer expert trajectories.

****

Local Calibration: Metrics and Recalibration
Rachel Luo,Aadyot Bhatnagar,Yu Bai,Shengjia Zhao,Huan Wang,Caiming Xiong,Silvio Savarese,Stefano Ermon,Edward Schmerling,Marco Pavone

Probabilistic classifiers output confidence scores along with their predictions, and these confidence scores should be calibrated, i.e., they should reflect the reliability of the prediction. Confidence scores that minimize standard metrics such as the expected calibration error (ECE) accurately measure the reliability on average across the entire population. However, it is in general impossible to measure the reliability of an individual prediction. In this work, we propose the local calibration error (LCE) to span the gap between average and individual reliability. For each individual prediction, the LCE measures the average reliability of a set of similar predictions, where similarity is quantified by a kernel function on a pretrained feature space and by a binning scheme over predicted model confidences. We show theoretically that the LCE can be estimated sample-efficiently from data, and empirically find that it reveals miscalibration modes that are more fine-grained than the ECE can detect. Our key result is a novel local recalibration method LoRe, to improve confidence scores for individual predictions and decrease the LCE. Experimentally, we show that our recalibration method produces more accurate confidence scores, which improves decision making and fairness on classification tasks using both image and tabular data.

****

Learning When and What to Ask: a Hierarchical Reinforcement Learning Framework
Khanh Xuan Nguyen,Yonatan Bisk,Hal Daumé III

Reliable AI agents should be mindful of the limits of their knowledge and consult humans when sensing that they do not have sufficient knowledge to make sound decisions. We formulate a hierarchical reinforcement learning framework for learning to decide when to request additional information from humans and what type of information would be helpful to request. Our framework extends partially-observed Markov decision processes (POMDPs) by allowing an agent to interact with an assistant to leverage their knowledge in accomplishing tasks. Results on a simulated human-assisted navigation problem demonstrate the effectiveness of our framework: aided with an interaction policy learned by our method, a navigation policy achieves up to a 7× improvement in task success rate compared to performing tasks only by itself. The interaction policy is also efficient: on average, only a quarter of all actions taken during a task execution are requests for informa tion. We analyze benefits and challenges of learning with a hierarchical policy structure and suggest directions for future work.

****

Towards Uncertainties in Deep Learning that Are Accurate and Calibrated

Volodymyr Kuleshov,Shachi Deshpande

Predictive uncertainties can be characterized by two properties---calibration and sharpness. This paper introduces algorithms that ensure the calibration of any model while maintaining sharpness. They apply in both classification and regression and guarantee the strong property of distribution calibration, while being simpler and more broadly applicable than previous methods (especially in the context of neural networks, which are often miscalibrated). Importantly, these algorithms achieve a long-standing statistical principle that forecasts should maximize sharpness subject to being fully calibrated. Using our algorithms, machine learning models can under some assumptions be calibrated without sacrificing accuracy: in a sense, calibration can be a free lunch. Empirically, we find that our methods improve predictive uncertainties on several tasks with minimal computational and implementation overhead.

****

Robust Models Are More Interpretable Because Attributions Look Normal

Zifan Wang,Matt Fredrikson,Anupam Datta

Recent work has found that adversarially-robust deep networks used for image classification are more interpretable: their feature attributions tend to be sharper, and are more concentrated on the objects associated with the image's ground-truth class. We show that smooth decision boundaries play an important role in this enhanced interpretability, as the model's input gradients around data points will more closely align with boundaries' normal vectors when they are smooth. Thus, because robust models have smoother boundaries, the results of gradient-based attribution methods will capture more accurate information about nearby decision boundaries. This understanding of robust interpretability leads to our second contribution: \emph{boundary attributions}, which aggregate information about the normal vectors of local decision boundaries to explain a classification outcome. We show that by leveraging the key factors underpinning robust interpretability, boundary attributions produce sharper, more concentrated visual explanations---even on non-robust models.

****

Trading Coverage for Precision: Conformal Prediction with Limited False Discoveries

Adam Fisch,Tal Schuster,Tommi S. Jaakkola,Regina Barzilay

In this paper, we develop a new approach to conformal prediction in which we aim to output a precise set of promising prediction candidates that is guaranteed to contain a limited number of incorrect answers. Standard conformal prediction provides the ability to adapt to model uncertainty by constructing a calibrated candidate set in place of a single prediction, with guarantees that the set contains the correct answer with high probability. In order to obey this coverage property, however, conformal sets can often become inundated with noisy candidates---which can render them unhelpful in practice. This is particularly relevant to large-scale settings where the cost (monetary or otherwise) of false positives is substantial, such as for in-silico screening for drug discovery, where any positively identified molecular compound is then manufactured and tested. We propose to trade coverage for precision by enforcing that the presence of incorrect candidates in the predicted conformal sets (i.e., the total number of false discoveries) is bounded according to a user-specified tolerance. Subject to this constraint, our algorithm then optimizes for a generalized notion of set coverage (i.e., the true discovery rate) that allows for any number of true answers for a given query (including zero). We demonstrate the effectiveness of this approach across a number of classification tasks in natural language processing, computer vision, and computational chemistry.

****

Communicating Natural Programs to Humans and Machines

Sam Acquaviva,Yewen Pu,Marta Kryven,Catherine Wong,Theodoros Sechopoulos,Gabrielle Ecanow,Maxwell Nye,Michael Henry Tessler,Joshua B. Tenenbaum

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 \emph{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.

****

On the benefits of maximum likelihood estimation for Regression and Forecasting

Pranjal Awasthi,Abhimanyu Das,Rajat Sen,Ananda Theertha Suresh

We advocate for a practical Maximum Likelihood Estimation (MLE) approach towards designing loss functions for regression and forecasting, as an alternative to the typical approach of direct empirical risk minimization on a specific target metric. The MLE approach is better suited to capture inductive biases such as prior domain knowledge in datasets, and can output post-hoc estimators at inference time that can optimize different types of target metrics. We present theoretical results to demonstrate that our approach is competitive with any estimator for the target metric under some general conditions. In two example practical settings, Poisson and Pareto regression, we show that our competitive results can be used to prove that the MLE approach has better excess risk bounds than directly minimizing the target metric. We also demonstrate empirically that our method instantiated with a well-designed general purpose mixture likelihood family can obtain superior performance for a variety of tasks across time-series forecasting and regression datasets with different data distributions.

****

AAVAE: Augmentation-Augmented Variational Autoencoders

William Alejandro Falcon,Ananya Harsh Jha,Teddy Koker,Kyunghyun Cho

Recent methods for self-supervised learning can be grouped into two paradigms: contrastive and non-contrastive approaches. Their success can largely be attributed to data augmentation pipelines which generate multiple views of a single input that preserve the underlying semantics. In this work, we introduce augmentation-augmented variational autoencoders (AAVAE), yet another alternative to self-supervised learning, based on autoencoding. We derive AAVAE starting from the conventional variational autoencoder (VAE), by replacing the KL divergence regularization, which is agnostic to the input domain, with data augmentations that explicitly encourage the internal representations to encode domain-specific invariances and equivariances. We empirically evaluate the proposed AAVAE on image classification, similar to how recent contrastive and non-contrastive learning algorithms have been evaluated. Our experiments confirm the effectiveness of data augmentation as a replacement for KL divergence regularization. The AAVAE outperforms the VAE by 30% on CIFAR-10, 40% on STL-10 and 45% on Imagenet. On CIFAR-10 and STL-10, the results for AAVAE are largely comparable to the state-of-the-art algorithms for self-supervised learning.

****

SVMnet: Non-parametric image classification based on convolutional SVM ensembles for small training sets

Hunter Goddard,Lior Shamir

Deep convolutional neural networks (DCNNs) have demonstrated superior power in their ability to classify image data. However, one of the downsides of DCNNs for supervised learning of image data is that their training normally requires large sets of labeled "ground truth" images. Since in many real-world problems large sets of pre-labeled images are not always available, DCNNs might not perform in an optimal manner in all real-world cases. Here we propose SVMnet -- a method based on a layered structure of Support Vector Machine (SVM) ensembles for non-parametric image classification. By utilizing the quick learning of SVMs compared to neural networks, the proposed method can reach higher accuracy than DCNNs when the training set is small. Experimental results show that while "conventional" DCNN architectures such as ResNet-50 outperform SVMnet when the size of the training set is large, SVMnet provides a much higher accuracy when the number of "ground truth" training samples is small.

****

Yformer: U-Net Inspired Transformer Architecture for Far Horizon Time Series Forecasting

Kiran Madhusudhanan,Johannes Burchert,Nghia Duong-Trung,Stefan Born,Lars Schmidt-Thieme

Time series data is ubiquitous in research as well as in a wide variety of industrial applications. Effectively analyzing the available historical data and providing insights into the far future allows us to make effective decisions. Recent research has witnessed the superior performance of transformer-based architectures, especially in the regime of far horizon time series forecasting. However, the current state of the art sparse Transformer architectures fail to couple down- and upsampling procedures to produce outputs in a similar resolution as the input. We propose the Yformer model, based on a novel Y-shaped encoder-decoder architecture that (1) uses direct connection from the downscaled encoder layer to the corresponding upsampled decoder layer in a U-Net inspired architecture, (2) Combines the downscaling/upsampling with sparse attention to capture long-range effects, and (3) stabilizes the encoder-decoder stacks with the addition of an auxiliary reconstruction loss. Extensive experiments have been conducted with relevant baselines on four benchmark datasets, demonstrating an average improvement of 19.82, 18.41 percentage MSE and 13.62, 11.85 percentage MAE in comparison to the current state of the art for the univariate and the multivariate settings respectively.

****

Finite-Time Error Bounds for Distributed Linear Stochastic Approximation

Yixuan Lin,Ji Liu,Vijay Gupta

This paper considers a novel multi-agent linear stochastic approximation algorithm driven by Markovian noise and general consensus-type interaction, in which each agent evolves according to its local stochastic approximation process which depends on the information from its neighbors. The interconnection structure among the agents is described by a time-varying directed graph. While the convergence of consensus-based stochastic approximation algorithms when the interconnection among the agents is described by doubly stochastic matrices (at least in expectation) has been studied, less is known about the case when the interconnection matrix is simply stochastic. For any uniformly strongly connected graph sequences whose associated interaction matrices are stochastic, the paper derives finite-time bounds on the mean-square error, defined as the deviation of the output of the algorithm from the unique equilibrium point of the associated ordinary differential equation. For the case of interconnection matrices being stochastic, the equilibrium point can be any unspecified convex combination of the local equilibria of all the agents in the absence of communication. Both the cases with constant and time-varying step-sizes are considered. In the case when the convex combination is required to be a straight average and interaction between any pair of neighboring agents may be uni-directional, so that doubly stochastic matrices cannot be implemented in a distributed manner, the paper proposes a push-type distributed stochastic approximation algorithm and provides its finite-time bound s for the performance by leveraging the analysis for the consensus-type algorithm with stochastic matrices.

****

Effect of scale on catastrophic forgetting in neural networks

Vinay Venkatesh Ramasesh,Aitor Lewkowycz,Ethan Dyer

Catastrophic forgetting presents a challenge in developing deep learning models capable of continual learning, i.e. learning tasks sequentially. Recently, both computer vision and natural-language processing have witnessed great progress through the use of large-scale pretrained models. In this work, we present an empirical study of catastrophic forgetting in this pretraining paradigm.

Our experiments indicate that large, pretrained ResNets and Transformers are significantly more resistant to forgetting than randomly-initialized, trained-from-scratch models; this robustness systematically improves with scale of both model and pretraining dataset size.

We take initial steps towards characterizing what aspect of model representations allows them to perform continual learning so well, finding that in the pretrained models, distinct class representations grow more orthogonal with scale. Our results suggest that, when possible, scale and a diverse pretraining dataset can be useful ingredients in mitigating catastrophic forgetting.

****

Pareto Navigation Gradient Descent: a First Order Algorithm for Optimization in Pareto Set

Mao Ye,qiang liu

Many modern machine learning applications, such as multi-task learning, require finding optimal model parameters to trade-off multiple objective functions that may conflict with each other.

The notion of the Pareto set allows us to focus on the set of (often infinite number of)

models that cannot be strictly improved. But it does not provide an actionable procedure for picking one or a few special models to return to practical users. In this paper, we consider \emph{optimization in Pareto set (OPT-in-Pareto)}, the problem of finding Pareto models that optimize an extra reference criterion function within the Pareto set. This function can either encode a specific preference from the users, or represent a generic diversity measure for obtaining a set of diversified Pareto models that are representative of the whole Pareto set.

Unfortunately, despite being a highly useful framework, efficient algorithms for OPT-in-Pareto have been largely missing, especially for large-scale, non-convex, and non-linear objectives in deep learning. A naive approach is to apply Riemannian manifold gradient descent on the Pareto set, which yields a high computational cost due to the need for eigen-calculation of Hessian matrices. We propose a first-order algorithm that approximately solves OPT-in-Pareto using only gradient information, with both high practical efficiency and theoretically guaranteed convergence property. Empirically, we demonstrate that our method works efficiently for a variety of challenging multi-task-related problems.

****

Federated Learning with Partial Model Personalization

Krishna Pillutla,Kshitiz Malik,Abdelrahman Mohamed,Michael Rabbat,Maziar Sanjabi,Lin Xiao

We propose and analyze a general framework of federated learning with partial model personalization. Compared with full model personalization, partial model personalization relies on domain knowledge to select a small portion of the model to personalize, thus imposing a much smaller on-device memory footprint. We propose two federated optimization algorithms for training partially personalized models, where the shared and personal parameters are updated either simultaneously or alternately on each device, but only the shared parameters are communicated and aggregated at the server. We give convergence analyses of both algorithms for minimizing smooth nonconvex functions, providing theoretical support of them for training deep learning models. Our experiments on real-world image and text datasets demonstrate that (a) partial model personalization can obtain most of the benefit of full model personalization with a small fraction of personalized pa rameters, and, (b) the alternating update algorithm often outperforms the simultaneous update algorithm.

****

Learn Locally, Correct Globally: A Distributed Algorithm for Training Graph Neural Networks

Morteza Ramezani,Weilin Cong,Mehrdad Mahdavi,Mahmut Kandemir,Anand Sivasubramaniam

Despite the recent success of Graph Neural Networks (GNNs), training GNNs on large graphs remains challenging. The limited resource capacities of the existing servers, the dependency between nodes in a graph, and the privacy concern due to the centralized storage and model learning have spurred the need to design an effective distributed algorithm for GNN training. However, existing distributed GNN training methods impose either excessive communication costs or large memory overheads that hinders their scalability. To overcome these issues, we propose a communication-efficient distributed GNN training technique named $\text{\textit{Learn Locally, Correct Globally}}$ (LLCG). To reduce the communication and memory overhead, each local machine in LLCG first trains a GNN on its local data by ignoring the dependency between nodes among different machines, then sends the locally trained model to the server for periodic model averaging. However, ignoring node dependency could result in significant performance degradation. To solve the performance degradation, we propose to apply $\text{\textit{Global Server Corrections}}$ on the server to refine the locally learned models. We rigorously analyze the convergence of distributed methods with periodic model averaging for training GNNs and show that naively applying periodic model averaging but ignoring the dependency between nodes will suffer from an irreducible residual error. However, this residual error can be eliminated by utilizing the proposed global corrections to entail fast convergence rate. Extensive experiments on real-world datasets show that LLCG can significantly improve the efficiency without hurting the performance.

****

Conditional Image Generation by Conditioning Variational Auto-Encoders

William Harvey,Saeid Naderiparizi,Frank Wood

We present a conditional variational auto-encoder (VAE) which, to avoid the substantial cost of training from scratch, uses an architecture and training objective capable of leveraging a foundation model in the form of a pretrained unconditional VAE. To train the conditional VAE, we only need to train an artifact to perform amortized inference over the unconditional VAE's latent variables given a conditioning input. We demonstrate our approach on tasks including image inpainting, for which it outperforms state-of-the-art GAN-based approaches at faithfully representing the inherent uncertainty. We conclude by describing a possible application of our inpainting model, in which it is used to perform Bayesian experimental design for the purpose of guiding a sensor.

****

Learning 3D Representations of Molecular Chirality with Invariance to Bond Rotations

Keir Adams,Lagnajit Pattanaik,Connor W. Coley

Molecular chirality, a form of stereochemistry most often describing relative spatial arrangements of bonded neighbors around tetrahedral carbon centers, influences the set of 3D conformers accessible to the molecule without changing its 2D graph connectivity. Chirality can strongly alter (bio)chemical interactions, particularly protein-drug binding. Most 2D graph neural networks (GNNs) designed for molecular property prediction at best use atomic labels to naïvely treat chirality, while E(3)-invariant 3D GNNs are invariant to chirality altogether. To enable representation learning on molecules with defined stereochemistry, we design an SE(3)-invariant model that processes torsion angles of a 3D molecular conformer. We explicitly model conformational flexibility by integrating a novel type of invariance to rotations about internal molecular bonds into the architecture, mitigating the need for multi-conformer data augmentation. We test our model on four benchmarks: contrastive learning to distinguish conformers of different stereoisomers in a learned latent space, classification of chiral centers as R/S, prediction of how enantiomers rotate circularly polarized light, and ranking enantiomers by their docking scores in an enantiosensitive protein pocket. We compare our model, Chiral InterRoto-Invariant Neural Network (ChIRo), with 2D and 3D GNNs to demonstrate that our model achieves state of the art performance when learning chiral-sensitive functions from molecular structures.

****

Neural Methods for Logical Reasoning over Knowledge Graphs

Alfonso Amayuelas,Shuai Zhang,Xi Susie Rao,Ce Zhang

Reasoning is a fundamental problem for computers and deeply studied in Artificial Intelligence. In this paper, we specifically focus on answering multi-hop logical queries on Knowledge Graphs (KGs). This is a complicated task because, in real world scenarios, the graphs tend to be large and incomplete. Most previous works have been unable to create models that accept full First-Order Logical (FOL) queries, which includes negative queries, and have only been able to process a limited set of query structures. Additionally, most methods present logic operators that can only perform the logical operation they are made for. We introduce a set of models that use Neural Networks to create one-point vector embeddings to answer the queries. The versatility of neural networks allows the framework to handle FOL queries with Conjunction, Disjunction and Negation operators. We demonstrate experimentally the performance of our models through extensive experimentation on well-known benchmarking datasets. Besides having more versatile operators, the models achieve a 10% relative increase over best performing state of the art and more than 30% over the original method based on single-point vector embeddings.

****

Consistent Counterfactuals for Deep Models

Emily Black,Zifan Wang,Matt Fredrikson

Counterfactual examples are one of the most commonly-cited methods for explaining the predictions of machine learning models in key areas such as finance and medical diagnosis. Counterfactuals are often discussed under the assumption that the model on which they will be used is static, but in deployment models may be periodically retrained or fine-tuned. This paper studies the consistency of model prediction on counterfactual examples in deep networks under small changes to initial training conditions, such as weight initialization and leave-one-out variations in data, as often occurs during model deployment. We demonstrate experimentally that counterfactual examples for deep models are often inconsistent across such small changes, and that increasing the cost of the counterfactual, a stability-enhancing mitigation suggested by prior work in the context of simpler models, is not a reliable heuristic in deep networks. Rather, our analysis shows that a model's Lipschitz continuity around the counterfactual, along with confidence of its prediction, is key to its consistency across related models. To this end, we propose Stable Neighbor Search as a way to generate more consistent counterfactual explanations, and illustrate the effectiveness of this approach on several benchmark datasets.

****

Centroid Approximation for Bootstrap

Mao Ye,qiang liu

Bootstrap is a principled and powerful frequentist statistical tool for uncertainty quantification. Unfortunately, standard bootstrap methods are computationally intensive due to the need of drawing a large i.i.d. bootstrap sample to approximate the ideal bootstrap distribution; this largely hinders their application in large-scale machine learning, especially deep learning problems. In this work, we propose an efficient method to explicitly \emph{optimize} a small set of high quality ``centroid'' points to better approximate the ideal bootstrap distribution. We achieve this by minimizing a simple objective function that is asymptotically equivalent to the Wasserstein distance to the ideal bootstrap distribution. This allows us to provide an accurate estimation of uncertainty with a small number of bootstrap centroids, outperforming the naive i.i.d. sampling approach. Empirically, we show that our method can boost the performance of bootstrap in a variety of applications.

****

Bounding Membership Inference

Anvith Thudi,I Shumailov,Franziska Boenisch,Nicolas Papernot

Differential Privacy (DP) is the de facto standard for reasoning about the privacy guarantees of a training algorithm. Despite the empirical observation that DP reduces the vulnerability of models to existing membership inference (MI) attacks, a theoretical underpinning as to why this is the case is largely missing in the literature. In practice, this means that models need to be trained with differential privacy guarantees that greatly decrease their accuracy. In this paper, we provide a tighter bound on the accuracy of any membership inference adversary when a training algorithm provides ϵ-DP. Our bound informs the design of a novel privacy amplification scheme, where an effective training set is sub-sampled from a larger set prior to the beginning of training, to greatly reduce the bound on MI accuracy. As a result, our scheme enables ϵ-DP users to employ looser differential privacy guarantees when training their model to limit the success of any MI adversary; this, in turn, ensures that the model's accuracy is less impacted by the privacy guarantee. Finally, we discuss the implications of our MI bound on machine unlearning.

****

Unified Visual Transformer Compression

Shixing Yu,Tianlong Chen,Jiayi Shen,Huan Yuan,Jianchao Tan,Sen Yang,Ji Liu,Zhangyang Wang

Vision transformers (ViTs) have gained popularity recently. Even without customized image operators such as convolutions, ViTs can yield competitive performance when properly trained on massive data. However, the computational overhead of ViTs remains prohibitive, due to stacking multi-head self-attention modules and else. Compared to the vast literature and prevailing success in compressing convolutional neural networks, the study of Vision Transformer compression has also just emerged, and existing works focused on one or two aspects of compression. This paper proposes a unified ViT compression framework that seamlessly assembles three effective techniques: pruning, layer skipping, and knowledge distillation. We formulate a budget-constrained, end-to-end optimization framework, targeting jointly learning model weights, layer-wise pruning ratios/masks, and skip configurations, under a distillation loss. The optimization problem is then solved using the primal-dual algorithm. Experiments are conducted with several ViT variants, e.g. DeiT and T2T-ViT backbones on the ImageNet dataset, and our approach consistently outperforms recent competitors. For example, DeiT-Tiny can be trimmed down to 50\% of the original FLOPs almost without losing accuracy. Codes are available online:~\url{https://github.com/VITA-Group/UVC}.

****

Training Multi-Layer Over-Parametrized Neural Network in Subquadratic Time

Zhao Song,Lichen Zhang,Ruizhe Zhang

In the recent years of development of theoretical machine learning, over-parametrization has been shown to be a powerful tool to resolve many fundamental problems, such as the convergence analysis of deep neural network. While many works have been focusing on designing various algorithms for over-parametrized network with one-hidden layer, multiple-hidden layers framework has received much less attention due to the complexity of the analysis, and even fewer algorithms have been proposed. In this work, we initiate the study of the performance of second-order algorithm on multiple-hidden layers over-parametrized neural network. We propose a novel algorithm to train such network, in time subquadratic in the width of the neural network. Our algorithm combines the Gram-Gauss-Newton method, tensor-based sketching techniques and preconditioning.

****

QTN-VQC: An End-to-End Learning Framework for Quantum Neural Networks

Jun Qi,Chao-Han Huck Yang,Pin-Yu Chen

The advent of noisy intermediate-scale quantum (NISQ) computers raises a crucial challenge to design quantum neural networks for fully quantum learning tasks. To bridge the gap, this work proposes an end-to-end learning framework named QTN-VQC, by introducing a trainable quantum tensor network (QTN) for quantum embeddi ng on a variational quantum circuit (VQC). The architecture of QTN is composed of a parametric tensor-train network for feature extraction and a tensor product encoding for quantum encoding. We highlight the QTN for quantum embedding in terms of two perspectives: (1) we theoretically characterize QTN by analyzing its representation power of input features; (2) QTN enables an end-to-end parametric model pipeline, namely QTN-VQC, from the generation of quantum embedding to the output measurement. Our experiments on the MNIST dataset demonstrate the advantages of QTN for quantum embedding over other quantum embedding approaches.

****

Half-Inverse Gradients for Physical Deep Learning

Patrick Schnell,Philipp Holl,Nils Thuerey

Recent works in deep learning have shown that integrating differentiable physics simulators into the training process can greatly improve the quality of results. Although this combination represents a more complex optimization task than usual neural network training, the same gradient-based optimizers are used to minimize the loss function. However, the integrated physics solvers have a profound effect on the gradient flow as manipulating scales in magnitude and direction is an inherent property of many physical processes. Consequently, the gradient flow is often highly unbalanced and creates an environment in which existing gradient-based optimizers perform poorly. In this work, we analyze the characteristics of both physical and neural network optimizations separately to derive a new method based on a half-inversion of the Jacobian. Our approach combines principles of both classical network and physics optimizers to solve the combined optimization task. Compared to state-of-the-art neural network optimizers, our method converges more quickly and to better solutions, which we demonstrate on three complex learning problems involving nonlinear oscillators, the Schroedinger equation and the Poisson problem.

****

Transformer-based Transform Coding

Yinhao Zhu,Yang Yang,Taco Cohen

Neural data compression based on nonlinear transform coding has made great progress over the last few years, mainly due to improvements in prior models, quantization methods and nonlinear transforms. A general trend in many recent works pushing the limit of rate-distortion performance is to use ever more expensive prior models that can lead to prohibitively slow decoding. Instead, we focus on more expressive transforms that result in a better rate-distortion-computation trade-off. Specifically, we show that nonlinear transforms built on Swin-transformers can achieve better compression efficiency than transforms built on convolutional neural networks (ConvNets), while requiring fewer parameters and shorter decoding time. Paired with a compute-efficient Channel-wise Auto-Regressive Model prior, our SwinT-ChARM model outperforms VTM-12.1 by 3.68% in BD-rate on Kodak with comparable decoding speed. In P-frame video compression setting, we are able to outperform the popular ConvNet-based scale-space-flow model by 12.35% in BD-rate on UVG. We provide model scaling studies to verify the computational efficiency of the proposed solutions and conduct several analyses to reveal the source of coding gain of transformers over ConvNets, including better spatial decorrelation, flexible effective receptive field, and more localized response of latent pixels during progressive decoding.

****

Object Pursuit: Building a Space of Objects via Discriminative Weight Generation

Chuanyu Pan,Yanchao Yang,Kaichun Mo,Yueqi Duan,Leonidas Guibas

We propose a framework to continuously learn object-centric representations for visual learning and understanding. Existing object-centric representations either rely on supervisions that individualize objects in the scene, or perform unsupervised disentanglement that can hardly deal with complex scenes in the real world. To mitigate the annotation burden and relax the constraints on the statistical complexity of the data, our method leverages interactions to effectively sample diverse variations of an object and the corresponding training signals while learning the object-centric representations. Throughout learning, objects are st reamed one by one in random order with unknown identities, and are associated with latent codes that can synthesize discriminative weights for each object through a convolutional hypernetwork. Moreover, re-identification of learned objects and forgetting prevention are employed to make the learning process efficient and robust. We perform an extensive study of the key features of the proposed framework and analyze the characteristics of the learned representations. Furthermore, we demonstrate the capability of the proposed framework in learning representations that can improve label efficiency in downstream tasks. Our code and trained models are made publicly available at: https://github.com/pptrick/Object-Pursuit.

****

Monotonic Improvement Guarantees under Non-stationarity for Decentralized PPO
Mingfei Sun,Sam Devlin,Jacob Austin Beck,Katja Hofmann,Shimon Whiteson

We present a new monotonic improvement guarantee for optimizing decentralized policies in cooperative Multi-Agent Reinforcement Learning (MARL), which holds even when the transition dynamics are non-stationary. This new analysis provides a theoretical understanding of the strong performance of two recent actor-critic methods for MARL, i.e., Independent Proximal Policy Optimization (IPPO) and Multi-Agent PPO (MAPPO), which both rely on independent ratios, i.e., computing probability ratios separately for each agent's policy. We show that, despite the non-stationarity that independent ratios cause, a monotonic improvement guarantee still arises as a result of enforcing the trust region constraint over joint policies. We also show this trust region constraint can be effectively enforced in a principled way by bounding independent ratios based on the number of agents in training, providing a theoretical foundation for proximal ratio clipping. Moreover, we show that the surrogate objectives optimized in IPPO and MAPPO are essentially equivalent when their critics converge to a fixed point. Finally, our empirical results support the hypothesis that the strong performance of IPPO and MAPPO is a direct result of enforcing such a trust region constraint via clipping in centralized training, and the good values of the hyperparameters for this enforcement are highly sensitive to the number of agents, as predicted by our theoretical analysis.

****

PAC Prediction Sets Under Covariate Shift
Sangdon Park,Edgar Dobriban,Insup Lee,Osbert Bastani

An important challenge facing modern machine learning is how to rigorously quantify the uncertainty of model predictions. Conveying uncertainty is especially important when there are changes to the underlying data distribution that might invalidate the predictive model. Yet, most existing uncertainty quantification algorithms break down in the presence of such shifts. We propose a novel approach that addresses this challenge by constructing \emph{probably approximately correct (PAC)} prediction sets in the presence of covariate shift. Our approach focuses on the setting where there is a covariate shift from the source distribution (where we have labeled training examples) to the target distribution (for which we want to quantify uncertainty). Our algorithm assumes given importance weights that encode how the probabilities of the training examples change under the covariate shift. In practice, importance weights typically need to be estimated; thus, we extend our algorithm to the setting where we are given confidence intervals for the importance weights. We demonstrate the effectiveness of our approach on covariate shifts based on DomainNet and ImageNet. Our algorithm satisfies the PAC constraint, and gives prediction sets with the smallest average normalized size among approaches that always satisfy the PAC constraint.

****

EE-Net: Exploitation-Exploration Neural Networks in Contextual Bandits
Yikun Ban,Yuchen Yan,Arindam Banerjee,Jingrui He

In this paper, we propose a novel neural exploration strategy in contextual bandits, EE-Net, distinct from the standard UCB-based and TS-based approaches. Contextual multi-armed bandits have been studied for decades with various applications. To solve the exploitation-exploration tradeoff in bandits, there are three main techniques: epsilon-greedy, Thompson Sampling (TS), and Upper Confidence Boun d (UCB). In recent literature, linear contextual bandits have adopted ridge regression to estimate the reward function and combine it with TS or UCB strategies for exploration. However, this line of works explicitly assumes the reward is based on a linear function of arm vectors, which may not be true in real-world datasets. To overcome this challenge, a series of neural bandit algorithms have been proposed, where a neural network is used to learn the underlying reward function and TS or UCB are adapted for exploration. Instead of calculating a large-deviation based statistical bound for exploration like previous methods, we propose "EE-Net", a novel neural-based exploration strategy. In addition to using a neural network (Exploitation network) to learn the reward function, EE-Net uses another neural network (Exploration network) to adaptively learn potential gains compared to the currently estimated reward for exploration. Then, a decision-maker is constructed to combine the outputs from the Exploitation and Exploration networks. We prove that EE-Net can achieve $\mathcal{O}(\sqrt{T\log T})$ regret and show that EE-Net outperforms existing linear and neural contextual bandit baselines on real-world datasets.

****

Generalization of Neural Combinatorial Solvers Through the Lens of Adversarial Robustness

Simon Geisler,Johanna Sommer,Jan Schuchardt,Aleksandar Bojchevski,Stephan Günnemann

End-to-end (geometric) deep learning has seen first successes in approximating the solution of combinatorial optimization problems. However, generating data in the realm of NP-hard/-complete tasks brings practical and theoretical challenges, resulting in evaluation protocols that are too optimistic. Specifically, most datasets only capture a simpler subproblem and likely suffer from spurious features. We investigate these effects by studying adversarial robustness -a local generalization property- to reveal hard, model-specific instances and spurious features. For this purpose, we derive perturbation models for SAT and TSP. Unlike in other applications, where perturbation models are designed around subjective notions of imperceptibility, our perturbation models are efficient and sound, allowing us to determine the true label of perturbed samples without a solver. Surprisingly, with such perturbations, a sufficiently expressive neural solver does not suffer from the limitations of the accuracy-robustness trade-off common in supervised learning. Although such robust solvers exist, we show empirically that the assessed neural solvers do not generalize well w.r.t. small perturbations of the problem instance.

****

One After Another: Learning Incremental Skills for a Changing World

Nur Muhammad Mahi Shafiullah,Lerrel Pinto

Reward-free, unsupervised discovery of skills is an attractive alternative to the bottleneck of hand-designing rewards in environments where task supervision is scarce or expensive. However, current skill pre-training methods, like many RL techniques, make a fundamental assumption -- stationary environments during training. Traditional methods learn all their skills simultaneously, which makes it difficult for them to both quickly adapt to changes in the environment, and to not forget earlier skills after such adaptation. On the other hand, in an evolving or expanding environment, skill learning must be able to adapt fast to new environment situations while not forgetting previously learned skills. These two conditions make it difficult for classic skill discovery to do well in an evolving environment. In this work, we propose a new framework for skill discovery, where skills are learned one after another in an incremental fashion. This framework allows newly learned skills to adapt to new environment or agent dynamics, while the fixed old skills ensure the agent doesn't forget a learned skill. We demonstrate experimentally that in both evolving and static environments, incremental skills significantly outperform current state-of-the-art skill discovery methods on both skill quality and the ability to solve downstream tasks. Videos for learned skills and code are made public on https://notmahi.github.io/disk

****

Takeuchi's Information Criteria as Generalization Measures for DNNs Close to NTK Regime

Hiroki Naganuma,Taiji Suzuki,Rio Yokota,Masahiro Nomura,Kohta Ishikawa,Ikuro Sato

Generalization measures are intensively studied in the machine learning community for better modeling generalization gaps. However, establishing a reliable generalization measure for statistical singular models such as deep neural networks (DNNs) is challenging due to the complex nature of the singular models.

We focus on a classical measure called Takeuchi's Information Criteria (TIC) to investigate allowed conditions in which the criteria can well explain generalization gaps caused by DNNs. In fact, theory indicates the applicability of TIC near the neural tangent kernel (NTK) regime.

Experimentally, we trained more than 5,000 DNN models with 12 DNN architectures including large models (e.g., VGG16) and 4 datasets, and estimated corresponding TICs in order to comprehensively study the relationship between the generalization gap and the TIC estimates.

We examine several approximation methods to estimate TIC with feasible computational load and investigate the accuracy trade-off. Experimental results indicate that estimated TIC well correlates generalization gaps under the conditions that are close to NTK regime. Outside the NTK regime, such correlation disappears, shown theoretically and empirically. We further demonstrate that TIC can yield better trial pruning ability for hyperparameter optimization over existing methods.

****

CheXT: Knowledge-Guided Cross-Attention Transformer for Abnormality Classification and Localization in Chest X-rays

Yan Han,Ying Ding,Ahmed Tewfik,Yifan Peng,Zhangyang Wang

Classical chest X-ray analysis has designed radiomic features to indicate the characteristics of abnormality of the chest X-rays. However, extracting reliable radiomic features heavily hinges on pathology localization, which is often absent in real-world image data. Although the past decade has witnessed the promising performance of convolutional neural networks (CNNs) in analyzing chest X-rays, most of them ignored domain knowledge such as radiomics. Recently, the surge of Transformers in computer vision has suggested a promising substitute for CNNs. It can encode highly expressive and generalizable representations and avoid costly manual annotations via a unique implementation of the self-attention mechanism. Moreover, Transformers naturally suit the feature extraction and fusion from different input modalities. Inspired by its recent success, this paper proposes \textbf{CheXT}, the first Transformer-based chest X-ray model. CheXT targets (semi-supervised) abnormality classification and localization from chest X-rays, enhanced by baked-in auxiliary knowledge guidance using radiomics. Specifically, CheXT consists of an image branch and a radiomics branch, interacted by cross-attention layers. During training, the image branch leverages its learned attention to estimate pathology localization, which is then utilized to extract radiomic features from images in the radiomics branch. Therefore, the two branches in CheXT are deeply fused and constitute an end-to-end optimization loop that can bootstrap accurate pathology localization from image data without any bounding box used for training. Extensive experiments on the NIH chest X-ray dataset demonstrate that CheXT significantly outperforms existing baselines in disease classification (by 1.1\% in average AUCs) and localization (by a \textbf{significant average margin of 3.6\%} over different IoU thresholds). Codes and models will be publicly released.

****

On-Target Adaptation

Dequan Wang,Shaoteng Liu,Sayna Ebrahimi,Evan Shelhamer,Trevor Darrell

Domain adaptation seeks to mitigate the shift between training on the source data and testing on the target data. Most adaptation methods rely on the source data by joint optimization over source and target. Source-free methods replace the source data with source parameters by fine-tuning the model on target. Either way, the majority of the parameter updates for the model representation and the cl assifier are derived from the source, and not the target. However, target accuracy is the goal, and so we argue for optimizing as much as possible on target. We show significant improvement by on-target adaptation, which learns the representation purely on target data, with only source predictions for supervision (without source data or parameter fine-tuning). In the long-tailed classification setting, we demonstrate on-target class distribution learning, which learns the (im)balance of classes on target data. On-target adaptation achieves state-of-the-art accuracy and computational efficiency on VisDA-C and ImageNet-Sketch. Learning more on target can deliver better models for target.

****

Graph-Guided Network for Irregularly Sampled Multivariate Time Series

Xiang Zhang,Marko Zeman,Theodoros Tsiligkaridis,Marinka Zitnik

In many domains, including healthcare, biology, and climate science, time series are irregularly sampled with varying time intervals between successive readouts and different subsets of variables (sensors) observed at different time points. Here, we introduce RAINDROP, a graph neural network that embeds irregularly sampled and multivariate time series while also learning the dynamics of sensors purely from observational data. RAINDROP represents every sample as a separate sensor graph and models time-varying dependencies between sensors with a novel message passing operator. It estimates the latent sensor graph structure and leverages the structure together with nearby observations to predict misaligned readouts. This model can be interpreted as a graph neural network that sends messages over graphs that are optimized for capturing time-varying dependencies among sensors. We use RAINDROP to classify time series and interpret temporal dynamics on three healthcare and human activity datasets. RAINDROP outperforms state-of-the-art methods by up to 11.4% (absolute F1-score points), including techniques that deal with irregular sampling using fixed discretization and set functions. RAINDROP shows superiority in diverse setups, including challenging leave-sensor-out settings.

****

FILM: Following Instructions in Language with Modular Methods

So Yeon Min,Devendra Singh Chaplot,Pradeep Kumar Ravikumar,Yonatan Bisk,Ruslan Salakhutdinov

Recent methods for embodied instruction following are typically trained end-to-end using imitation learning. This often requires the use of expert trajectories and low-level language instructions. Such approaches assume that neural states will integrate multimodal semantics to perform state tracking, building spatial memory, exploration, and long-term planning. In contrast, we propose a modular method with structured representations that (1) builds a semantic map of the scene and (2) performs exploration with a semantic search policy, to achieve the natural language goal. Our modular method achieves SOTA performance (24.46 %) with a substantial (8.17 % absolute) gap from previous work while using less data by eschewing both expert trajectories and low-level instructions. Leveraging low-level language, however, can further increase our performance (26.49 %). Our findings suggest that an explicit spatial memory and a semantic search policy can provide a stronger and more general representation for state-tracking and guidance, even in the absence of expert trajectories or low-level instructions.

****

The Evolution of Uncertainty of Learning in Games

Yun Kuen Cheung,Georgios Piliouras,Yixin Tao

Learning in games has become an object of intense interest for ML due to its connections to numerous AI architectures. We study standard online learning in games but from a non-standard perspective. Instead of studying the behavior of a single initial condition and whether it converges to equilibrium or not, we study the behavior of a probability distribution/measure over a set of initial conditions. This initial uncertainty is well-motivated both from a standard game-theoretic perspective (e.g. a modeler's uncertainty about the agents' initial beliefs) as well as from a ML one (e.g. noisy measurements, system initialization from a dataset distribution). Despite this, little is formally known about whether and under what conditions uncertainty is amplified or reduced in these systems. We use the popular measure of differential entropy to quantify the evolution of uncertainty. We find that such analysis shares an intimate relationship with volume analysis, a technique which was recently used to demonstrate the occurrence of Lyapunov chaos when using Multiplicative Weights Update (MWU) or Follow-the-Regularized-Leader (FTRL) algorithms in zero-sum games. This allows us to show that the differential entropy of these learning-in-game systems increases linearly with time, formalizing their increased unpredictability over time. We showcase the power of the framework by applying it in the study of multiple related systems, including different standard online optimization algorithms in numerous games and dynamics of evolutionary game theory.

****

Model Validation Using Mutated Training Labels: An Exploratory Study

Jie Zhang,Mark Harman,Benjamin Guedj,Earl Barr,John Shawe-Taylor

For out-of-sample validation, the sample set may be too small to be representative of the data distribution; the accuracy can have a large variance across different runs; excessive reuse of a fixed set of samples can lead to overfitting even if the samples are held out and not used in the training process. This paper introduces an exploratory study on Mutation Validation (MV), a model validation method using mutated training labels for supervised learning. MV mutates training data labels, retrains the model against the mutated data, then uses the metamorphic relation capturing the consequent training performance changes to assess model fit. It uses neither validation nor test set. The intuition underpinning MV is that overfitted models tend to fit noise in the training data. We explore 8 different learning algorithms, 18 datasets, and 5 types of hyperparameter tuning tasks. Our results demonstrate that MV is accurate in model selection: the model recommendation hit rate is 92% for MV and less than 60% for out-of-sample validation. MV also provides more stable hyperparameter tuning results than out-of-sample validation across different runs.

****

Pruning Edges and Gradients to Learn Hypergraphs from Larger Sets

David W Zhang,Gertjan J. Burghouts,Cees G. M. Snoek

This paper aims for set-to-hypergraph prediction, where the goal is to infer the set of relations for a given set of entities. This is a common abstraction for applications in particle physics, biological systems and combinatorial optimization. We address two common scaling problems encountered in set-to-hypergraph tasks that limit the size of the input set: the exponentially growing number of hyperedges and the run-time complexity, both leading to higher memory requirements. We make three contributions. First, we propose to predict and supervise the \emph{positive} edges only, which changes the asymptotic memory scaling from exponential to linear. Second, we introduce a training method that encourages iterative refinement of the predicted hypergraph, which allows us to skip iterations in the backward pass for improved efficiency and constant memory usage. Third, we combine both contributions in a single set-to-hypergraph model that enables us to address problems with larger input set sizes. We provide ablations for our main technical contributions and show that our model outperforms prior state-of-the-art, especially for larger sets.

****

Spike-inspired rank coding for fast and accurate recurrent neural networks

Alan Jeffares,Qinghai Guo,Pontus Stenetorp,Timoleon Moraitis

Biological spiking neural networks (SNNs) can temporally encode information in their outputs, e.g. in the rank order in which neurons fire, whereas artificial neural networks (ANNs) conventionally do not. As a result, models of SNNs for neuromorphic computing are regarded as potentially more rapid and efficient than ANNs when dealing with temporal input. On the other hand, ANNs are simpler to train, and usually achieve superior performance. Here we show that temporal coding such as rank coding (RC) inspired by SNNs can also be applied to conventional ANNs such as LSTMs, and leads to computational savings and speedups.

In our RC for ANNs, we apply backpropagation through time using the standard real-valued activations, but only from a strategically early time step of each sequ ential input example, decided by a threshold-crossing event. Learning then incorporates naturally also when to produce an output, without other changes to the model or the algorithm. Both the forward and the backward training pass can be significantly shortened by skipping the remaining input sequence after that first event. RC-training also significantly reduces time-to-insight during inference, with a minimal decrease in accuracy. The desired speed-accuracy trade-off is tunable by varying the threshold or a regularization parameter that rewards output entropy. We demonstrate these in two toy problems of sequence classification, and in a temporally-encoded MNIST dataset where our RC model achieves 99.19% accuracy after the first input time-step, outperforming the state of the art in temporal coding with SNNs, as well as in spoken-word classification of Google Speech Commands, outperforming non-RC-trained early inference with LSTMs.

****

Einops: Clear and Reliable Tensor Manipulations with Einstein-like Notation

Alex Rogozhnikov

Tensor computations underlie modern scientific computing and deep learning.
A number of tensor frameworks emerged varying in execution model, hardware support, memory management, model definition, etc.
However, tensor operations in all frameworks follow the same paradigm.
Recent neural network architectures demonstrate demand for higher expressiveness of tensor operations.
The current paradigm is not suited to write readable, reliable, or easy-to-modify code for multidimensional tensor manipulations.
Moreover, some commonly used operations do not provide sufficient checks and can break a tensor structure.
These mistakes are elusive as no tools or tests can detect them.
Independently, API discrepancies complicate code transfer between frameworks.
We propose einops notation: a uniform and generic way to manipulate tensor structure, that significantly improves code readability and flexibility by focusing on the structure of input and output tensors.
We implement einops notation in a Python package that efficiently supports multiple widely used frameworks and provides framework-independent minimalist API for tensor manipulations.

****

Explainable GNN-Based Models over Knowledge Graphs

David Jaime Tena Cucala,Bernardo Cuenca Grau,Egor V. Kostylev,Boris Motik

Graph Neural Networks (GNNs) are often used to learn transformations of graph data. While effective in practice, such approaches make predictions via numeric manipulations so their output cannot be easily explained symbolically. We propose a new family of GNN-based transformations of graph data that can be trained effectively, but where all predictions can be explained symbolically as logical inferences in Datalog—a well-known rule-based formalism. In particular, we show how to encode an input knowledge graph into a graph with numeric feature vectors, process this graph using a GNN, and decode the result into an output knowledge graph. We use a new class of monotonic GNNs (MGNNs) to ensure that this process is equivalent to a round of application of a set of Datalog rules. We also show that, given an arbitrary MGNN, we can automatically extract rules that completely characterise the transformation. We evaluate our approach by applying it to classification tasks in knowledge graph completion.

****

Language Modeling using LMUs: 10x Better Data Efficiency or Improved Scaling Compared to Transformers

Narsimha Reddy Chilkuri,Eric Hunsberger,Aaron Russell Voelker,Gurshaant Singh Malik,Chris Eliasmith

Recent studies have demonstrated that the performance of transformers on the task of language modeling obeys a power-law relationship with model size over six orders of magnitude. While transformers exhibit impressive scaling, their performance hinges on processing large amounts of data, and their computational and memory requirements grow quadratically with sequence length. Motivated by these considerations, we construct a Legendre Memory Unit based model that introduces a g eneral prior for sequence processing and exhibits an $O(n)$ and $O(n \ln n)$ (or better) dependency for memory and computation respectively. Over three orders of magnitude, we show that our new architecture attains the same accuracy as transformers with 10x fewer tokens. We also show that for the same amount of training our model improves the loss over transformers about as much as transformers improve over LSTMs. Additionally, we demonstrate that adding global self-attention complements our architecture and the augmented model improves performance even further.

****

How to Robustify Black-Box ML Models? A Zeroth-Order Optimization Perspective

Yimeng Zhang,Yuguang Yao,Jinghan Jia,Jinfeng Yi,Mingyi Hong,Shiyu Chang,Sijia Liu

The lack of adversarial robustness has been recognized as an important issue for state-of-the-art machine learning (ML) models, e.g., deep neural networks (DNNs). Thereby, robustifying ML models against adversarial attacks is now a major focus of research. However, nearly all existing defense methods, particularly for robust training, made the white-box assumption that the defender has the access to the details of an ML model (or its surrogate alternatives if available), e.g., its architectures and parameters. Beyond existing works, in this paper we aim to address the problem of black-box defense: How to robustify a black-box model using just input queries and output feedback? Such a problem arises in practical scenarios, where the owner of the predictive model is reluctant to share model information in order to preserve privacy. To this end, we propose a general notion of defensive operation that can be applied to black-box models, and design it through the lens of denoised smoothing (DS), a first-order (FO) certified defense technique. To allow the design of merely using model queries, we further integrate DS with the zeroth-order (gradient-free) optimization. However, a direct implementation of zeroth-order (ZO) optimization suffers a high variance of gradient estimates, and thus leads to ineffective defense. To tackle this problem, we next propose to prepend an autoencoder (AE) to a given (black-box) model so that DS can be trained using variance-reduced ZO optimization. We term the eventual defense as ZO-AE-DS. In practice, we empirically show that ZO-AE-DS can achieve improved accuracy, certified robustness, and query complexity over existing baselines. And the effectiveness of our approach is justified under both image classification and image reconstruction tasks.

****

Mention Memory: incorporating textual knowledge into Transformers through entity mention attention

Michiel de Jong,Yury Zemlyanskiy,Nicholas FitzGerald,Fei Sha,William W. Cohen

Natural language understanding tasks such as open-domain question answering often require retrieving and assimilating factual information from multiple sources. We propose to address this problem by integrating a semi-parametric representation of a large text corpus into a Transformer model as a source of factual knowledge.

Specifically, our method represents knowledge with ``mention memory'', a table of dense vector representations of every entity mention in a corpus. The proposed model - TOME - is a Transformer that accesses the information through internal memory layers in which each entity mention in the input passage attends to the mention memory. This approach enables synthesis of and reasoning over many disparate sources of information within a single Transformer model.

In experiments using a memory of 150 million Wikipedia mentions, TOME achieves strong performance on several open-domain knowledge-intensive tasks, including the claim verification benchmarks HoVer and FEVER and several entity-based QA benchmarks. We also show that the model learns to attend to informative mentions without any direct supervision. Finally we demonstrate that the model can generalize to new unseen entities by updating the memory without retraining.

****

Training Data Generating Networks: Shape Reconstruction via Bi-level Optimization

Biao Zhang,Peter Wonka

We propose a novel 3d shape representation for 3d shape reconstruction from a single image. Rather than predicting a shape directly, we train a network to generate a training set which will be fed into another learning algorithm to define the shape. The nested optimization problem can be modeled by bi-level optimization. Specifically, the algorithms for bi-level optimization are also being used in meta learning approaches for few-shot learning. Our framework establishes a link between 3D shape analysis and few-shot learning. We combine training data generating networks with bi-level optimization algorithms to obtain a complete framework for which all components can be jointly trained. We improve upon recent work on standard benchmarks for 3d shape reconstruction.

****

Monotonic Differentiable Sorting Networks

Felix Petersen,Christian Borgelt,Hilde Kuehne,Oliver Deussen

Differentiable sorting algorithms allow training with sorting and ranking supervision, where only the ordering or ranking of samples is known. Various methods have been proposed to address this challenge, ranging from optimal transport-based differentiable Sinkhorn sorting algorithms to making classic sorting networks differentiable. One problem of current differentiable sorting methods is that they are non-monotonic. To address this issue, we propose a novel relaxation of conditional swap operations that guarantees monotonicity in differentiable sorting networks. We introduce a family of sigmoid functions and prove that they produce differentiable sorting networks that are monotonic. Monotonicity ensures that the gradients always have the correct sign, which is an advantage in gradient-based optimization. We demonstrate that monotonic differentiable sorting networks improve upon previous differentiable sorting methods.

****

CrowdPlay: Crowdsourcing Human Demonstrations for Offline Learning

Matthias Gerstgrasser,Rakshit Trivedi,David C. Parkes

Crowdsourcing has been instrumental for driving AI advances that rely on large-scale data. At the same time, reinforcement learning has seen rapid progress through benchmark environments that strike a balance between tractability and real-world complexity, such as ALE and OpenAI Gym. In this paper, we aim to fill a gap at the intersection of these two: The use of crowdsourcing to generate large-scale human demonstration data in the support of advancing research into imitation learning and offline learning.

To this end, we present CrowdPlay, a complete crowdsourcing pipeline for any standard RL environment including OpenAI Gym (made available under an open-source license); a large-scale publicly available crowdsourced dataset of human gameplay demonstrations in Atari 2600 games, including multimodal behavior and human-human and human-AI multiagent data; offline learning benchmarks with extensive human data evaluation; and a detailed study of incentives, including real-time feedback to drive high quality data.

We hope that this will drive the improvement in design of algorithms that account for the complexity of human, behavioral data and thereby enable a step forward in direction of effective learning for real-world settings. Our code and dataset are available at https://mgerstgrasser.github.io/crowdplay/.

****

Soft Actor-Critic with Inhibitory Networks for Faster Retraining

Jaime S. Ide,Daria Micovic,Adrian P Pope,Michael John Guarino,Kevin Alcedo,David Rosenbluth

Reusing previously trained models is critical in deep reinforcement learning to speed up training of new agents. However, it is unclear how to acquire new skills when objectives and constraints are in conflict with previously learned skills. Moreover, when retraining, there is an intrinsic conflict between exploiting what has already been learned and exploring new skills. In soft actor-critic (SAC) methods, a temperature parameter can be dynamically adjusted to weight the action entropy and balance the explore $\times$ exploit trade-off. However, controlling a single coefficient can be challenging within the context of retraining, even more so when goals are contradictory. In this work, inspired by neuroscience research, we propose a novel approach using inhibitory networks to allow separa te and adaptive state value evaluations, as well as distinct automatic entropy tuning. Ultimately, our approach allows for controlling inhibition to handle conflict between exploiting less risky, acquired behaviors and exploring novel ones to overcome more challenging tasks. We validate our method through experiments in OpenAI Gym environments.

****

Model Agnostic Interpretability for Multiple Instance Learning

Joseph Early,Christine Evers,SArvapali Ramchurn

In Multiple Instance Learning (MIL), models are trained using bags of instances, where only a single label is provided for each bag. A bag label is often only determined by a handful of key instances within a bag, making it difficult to interpret what information a classifier is using to make decisions. In this work, we establish the key requirements for interpreting MIL models. We then go on to develop several model-agnostic approaches that meet these requirements. Our methods are compared against existing inherently interpretable MIL models on several datasets, and achieve an increase in interpretability accuracy of up to 30%. We also examine the ability of the methods to identify interactions between instances and scale to larger datasets, improving their applicability to real-world problems.

****

RelaxLoss: Defending Membership Inference Attacks without Losing Utility

Dingfan Chen,Ning Yu,Mario Fritz

As a long-term threat to the privacy of training data, membership inference attacks (MIAs) emerge ubiquitously in machine learning models.

Existing works evidence strong connection between the distinguishability of the training and testing loss distributions and the model's vulnerability to MIAs. Motivated by existing results, we propose a novel training framework based on a relaxed loss ($\textbf{RelaxLoss}$) with a more achievable learning target, which leads to narrowed generalization gap and reduced privacy leakage. RelaxLoss is applicable to any classification model with added benefits of easy implementation and negligible overhead. Through extensive evaluations on five datasets with diverse modalities (images, medical data, transaction records), our approach consistently outperforms state-of-the-art defense mechanisms in terms of resilience against MIAs as well as model utility. Our defense is the first that can withstand a wide range of attacks while preserving (or even improving) the target model's utility.

****

FastSHAP: Real-Time Shapley Value Estimation

Neil Jethani,Mukund Sudarshan,Ian Connick Covert,Su-In Lee,Rajesh Ranganath

Although Shapley values are theoretically appealing for explaining black-box models, they are costly to calculate and thus impractical in settings that involve large, high-dimensional models. To remedy this issue, we introduce FastSHAP, a new method for estimating Shapley values in a single forward pass using a learned explainer model. To enable efficient training without requiring ground truth Shapley values, we develop an approach to train FastSHAP via stochastic gradient descent using a weighted least-squares objective function. In our experiments with tabular and image datasets, we compare FastSHAP to existing estimation approaches and find that it generates accurate explanations with an orders-of-magnitude speedup.

****

Continual Backprop: Stochastic Gradient Descent with Persistent Randomness

Shibhansh Dohare,Richard S. Sutton,A. Rupam Mahmood

The Backprop algorithm for learning in neural networks utilizes two mechanisms: first, stochastic gradient descent and second, initialization with small random weights, where the latter is essential to the effectiveness of the former. We show that in continual learning setups, Backprop performs well initially, but over time its performance degrades. Stochastic gradient descent alone is insufficient to learn continually; the initial randomness enables only initial learning but not continual learning. To the best of our knowledge, ours is the first result showing this degradation in Backprop’s ability to learn. To address this issue,

we propose an algorithm that continually injects random features alongside gradient descent using a new generate-and-test process. We call this the Continual Backprop algorithm. We show that, unlike Backprop, Continual Backprop is able to continually adapt in both supervised and reinforcement learning problems. We expect that as continual learning becomes more common in future applications, a method like Continual Backprop will be essential where the advantages of random initialization are present throughout learning.
****
Model-Efficient Deep Learning with Kernelized Classification
Sadeep Jayasumana,Srikumar Ramalingam,Sanjiv Kumar
We investigate the possibility of using the embeddings produced by a lightweight network more effectively with a nonlinear classification layer. Although conventional deep networks use an abundance of nonlinearity for representation (embedding) learning, they almost universally use a linear classifier on the learned embeddings. This is suboptimal since better nonlinear classifiers could exist in the same embedding vector space. We advocate a nonlinear kernelized classification layer for deep networks to tackle this problem. We theoretically show that our classification layer optimizes over all possible kernel functions on the space of embeddings to learn an optimal nonlinear classifier. We then demonstrate the usefulness of this layer in learning more model-efficient classifiers in a number of computer vision and natural language processing tasks.
****
When, Why, and Which Pretrained GANs Are Useful?
Timofey Grigoryev,Andrey Voynov,Artem Babenko
The literature has proposed several methods to finetune pretrained GANs on new datasets, which typically results in higher performance compared to training from scratch, especially in the limited-data regime. However, despite the apparent empirical benefits of GAN pretraining, its inner mechanisms were not analyzed in-depth, and understanding of its role is not entirely clear. Moreover, the essential practical details, e.g., selecting a proper pretrained GAN checkpoint, currently do not have rigorous grounding and are typically determined by trial and error.

This work aims to dissect the process of GAN finetuning. First, we show that initializing the GAN training process by a pretrained checkpoint primarily affects the model's coverage rather than the fidelity of individual samples. Second, we explicitly describe how pretrained generators and discriminators contribute to the finetuning process and explain the previous evidence on the importance of pretraining both of them. Finally, as an immediate practical benefit of our analysis, we describe a simple recipe to choose an appropriate GAN checkpoint that is the most suitable for finetuning to a particular target task. Importantly, for most of the target tasks, Imagenet-pretrained GAN, despite having poor visual quality, appears to be an excellent starting point for finetuning, resembling the typical pretraining scenario of discriminative computer vision models.
****
A global convergence theory for deep ReLU implicit networks via over-parameterization
Tianxiang Gao,Hailiang Liu,Jia Liu,Hridesh Rajan,Hongyang Gao
Implicit deep learning has received increasing attention recently due to the fact that it generalizes the recursive prediction rule of many commonly used neural network architectures. Its prediction rule is provided implicitly based on the solution of an equilibrium equation. Although a line of recent empirical studies has demonstrated its superior performances, the theoretical understanding of implicit neural networks is limited. In general, the equilibrium equation may not be well-posed during the training. As a result, there is no guarantee that a vanilla (stochastic) gradient descent (SGD) training nonlinear implicit neural networks can converge. This paper fills the gap by analyzing the gradient flow of Rectified Linear Unit (ReLU) activated implicit neural networks. For an m width implicit neural network with ReLU activation and n training samples, we show that a randomly initialized gradient descent converges to a global minimum at a l

inear rate for the square loss function if the implicit neural network is over-parameterized. It is worth noting that, unlike existing works on the convergence of (S)GD on finite-layer over-parameterized neural networks, our convergence results hold for implicit neural networks, where the number of layers is infinite.
****

CausalDyna: Improving Generalization of Dyna-style Reinforcement Learning via Counterfactual-Based Data Augmentation
Deyao Zhu,Li Erran Li,Mohamed Elhoseiny

Deep reinforcement learning agents trained in real-world environments with a limited diversity of object properties to learn manipulation tasks tend to suffer overfitting and fail to generalize to unseen testing environments. To improve the agents' ability to generalize to object properties rarely seen or unseen, we propose a data-efficient reinforcement learning algorithm, CausalDyna, that exploits structural causal models (SCMs) to model the state dynamics. The learned SCM enables us to counterfactually reason what would have happened had the object had a different property value. This can help remedy limitations of real-world environments or avoid risky exploration of robots (e.g., heavy objects may damage the robot). We evaluate our algorithm in the CausalWorld robotic-manipulation environment. When augmented with counterfactual data, our CausalDyna outperforms state-of-the-art model-based algorithm, MBPO and model-free algorithm, SAC in both sample efficiency by up to 17% and generalization by up to 30%. Code will be made publicly available.
****

Learnability Lock: Authorized Learnability Control Through Adversarial Invertible Transformations
Weiqi Peng,Jinghui Chen

Owing much to the revolution of information technology, recent progress of deep learning benefits incredibly from the vastly enhanced access to data available in various digital formats. Yet those publicly accessible information also raises a fundamental issue concerning Intellectual Property, that is, how to precisely control legal or illegal exploitation of a dataset for training commercial models. To tackle this issue, this paper introduces and investigates a new concept called ''learnability lock'' for securing the process of data authorization. In particular, we propose adversarial invertible transformation, that can be viewed as a mapping from image to image, to encrypt data samples so that they become ''unlearnable'' by machine learning models with negligible loss of visual features. Meanwhile, authorized clients can use a specific key to unlock the learnability of the protected dataset and train models normally. The proposed learnability lock leverages class-wise perturbation that applies a universal transformation function on data samples of the same label. This ensures that the learnability can be easily restored with a simple inverse transformation while remaining difficult to be detected or reverse-engineered. We empirically demonstrate the success and practicability of our method on visual classification tasks.
****

Safe Deep RL in 3D Environments using Human Feedback
Matthew Rahtz,Vikrant Varma,Ramana Kumar,Zachary Kenton,Shane Legg,Jan Leike

Agents should avoid unsafe behaviour during both training and deployment. This typically requires a simulator and a procedural specification of unsafe behaviour. Unfortunately, a simulator is not always available, and procedurally specifying constraints can be difficult or impossible for many real-world tasks. A recently introduced technique, ReQueST, aims to solve this problem by learning a neural simulator of the environment from safe human trajectories, then using the learned simulator to efficiently learn a reward model from human feedback. However, it is yet unknown whether this approach is feasible in complex 3D environments with feedback obtained from real humans - whether sufficient pixel-based neural simulator quality can be achieved, and whether the human data requirements are viable in terms of both quantity and quality. In this paper we answer this question in the affirmative, using ReQueST to train an agent to perform a 3D first-person object collection task using data entirely from human contractors. We show that the resulting agent exhibits an order of magnitude reduction in unsafe behavi

our compared to standard reinforcement learning.
****

Learning to Infer the Structure of Network Games

Emanuele Rossi,Federico Monti,Yan Leng,Michael M. Bronstein,Xiaowen Dong

Strategic interactions between a group of individuals or organisations can be modelled as games played on networks, where a player's payoff depends not only on their actions but also on those of their neighbors.

Inferring the network structure from observed game outcomes (equilibrium actions) is an important problem with numerous potential applications in economics and social sciences.

Currently available methods require the knowledge of the utility function associated with the game, which is often unrealistic to obtain in real-world scenarios. To address this limitation, we propose a novel transformer-like architecture which correctly accounts for the symmetries of the problem and learns a mapping from the equilibrium actions to the network structure of the game without explicit knowledge of the utility function. We test our method on three different types of network games using both synthetic and real-world data, and demonstrate its effectiveness in network structure inference and superior performance over existing methods.
****

Private Multi-Task Learning: Formulation and Applications to Federated Learning

Shengyuan Hu,Steven Wu,Virginia Smith

Many problems in machine learning rely on multi-task learning (MTL), in which the goal is to solve multiple related machine learning tasks simultaneously. MTL is particularly relevant for privacy-sensitive applications in areas such as healthcare, finance, and IoT computing, where sensitive data from multiple, varied sources are shared for the purpose of learning. In this work, we formalize notions of task-level privacy for MTL via joint differential privacy(JDP), a relaxation of differential privacy for mechanism design and distributed optimization. We then propose an algorithm for mean-regularized MTL, an objective commonly used for applications in personalized federated learning, subject to JDP. We analyze our objective and solver, providing certifiable guarantees on both privacy and utility. Empirically, we find that our method allows for improved privacy/utility trade-offs relative to global baselines across common federated learning benchmarks.
****

On the Unreasonable Effectiveness of Feature Propagation in Learning on Graphs with Missing Node Features

Emanuele Rossi,Henry Kenlay,Maria I. Gorinova,Benjamin Paul Chamberlain,Xiaowen Dong,Michael M. Bronstein

While Graph Neural Networks (GNNs) have recently become the {\em de facto} standard for modeling relational data, they impose a strong assumption on the availability of the node or edge features of the graph. In many real-world applications, however,

features are only partially available; for example, in social networks, age and gender are available only for a small subset of users. We present a general approach for handling missing features in graph machine learning applications that is based on minimizing the Dirichlet energy and leads to a diffusion-type differential equation on the graph. The discretization of this equation produces a simple, fast and scalable algorithm which we call Feature Propagation. We experimentally show that the proposed approach outperforms previous methods on six common node-classification benchmarks and can withstand surprisingly high rates of missing features: on average we observe only around 5% relative accuracy drop when 99% of the features are missing. Moreover, it takes only 10 seconds to run on a graph with ~2.5M nodes and ~123M edges on a single GPU.
****

Grounding Aleatoric Uncertainty in Unsupervised Environment Design

Minqi Jiang,Michael D Dennis,Jack Parker-Holder,Andrei Lupu,Heinrich Kuttler,Edward Grefenstette,Tim Rocktäschel,Jakob Nicolaus Foerster

In reinforcement learning (RL), adaptive curricula have proven highly effective for learning policies that generalize well under a wide variety of changes to the environment. Recently, the framework of Unsupervised Environment Design (UED) generalized notions of curricula for RL in terms of generating entire environments, leading to the development of new methods with robust minimax-regret properties. However, in partially-observable or stochastic settings (those featuring aleatoric uncertainty), optimal policies may depend on the ground-truth distribution over the aleatoric features of the environment. Such settings are potentially problematic for curriculum learning, which necessarily shifts the environment distribution used during training with respect to the fixed ground-truth distribution in the intended deployment environment. We formalize this phenomenon as curriculum-induced covariate shift, and describe how, when the distribution shift occurs over such aleatoric environment parameters, it can lead to learning suboptimal policies. We then propose a method which, given black box access to a simulator, corrects this resultant bias by aligning the advantage estimates to the ground-truth distribution over aleatoric parameters. This approach leads to a minimax-regret UED method, SAMPLR, with Bayes-optimal guarantees.

****

Analyzing and Improving the Optimization Landscape of Noise-Contrastive Estimation

Bingbin Liu,Elan Rosenfeld,Pradeep Kumar Ravikumar,Andrej Risteski

Noise-contrastive estimation (NCE) is a statistically consistent method for learning unnormalized probabilistic models. It has been empirically observed that the choice of the noise distribution is crucial for NCE’s performance. However, such observation has never been made formal or quantitative. In fact, it is not even clear whether the difficulties arising from a poorly chosen noise distribution are statistical or algorithmic in nature.

In this work, we formally pinpoint reasons for NCE’s poor performance when an inappropriate noise distribution is used. Namely, we prove these challenges arise due to an ill-behaved (more precisely, flat) loss landscape.

To address this, we introduce a variant of NCE called \emph{eNCE} which uses an exponential loss and for which \emph{normalized gradient descent} addresses the landscape issues \emph{provably} when the target and noise distributions are in a given exponential family.

****

Federated Learning from Only Unlabeled Data with Class-conditional-sharing Clients

Nan Lu,Zhao Wang,Xiaoxiao Li,Gang Niu,Qi Dou,Masashi Sugiyama

Supervised federated learning (FL) enables multiple clients to share the trained model without sharing their labeled data. However, potential clients might even be reluctant to label their own data, which could limit the applicability of FL in practice. In this paper, we show the possibility of unsupervised FL whose model is still a classifier for predicting class labels, if the class-prior probabilities are shifted while the class-conditional distributions are shared among the unlabeled data owned by the clients. We propose federation of unsupervised learning (FedUL), where the unlabeled data are transformed into surrogate labeled data for each of the clients, a modified model is trained by supervised FL, and the wanted model is recovered from the modified model. FedUL is a very general solution to unsupervised FL: it is compatible with many supervised FL methods, and the recovery of the wanted model can be theoretically guaranteed as if the data have been labeled. Experiments on benchmark and real-world datasets demonstrate the effectiveness of FedUL. Code is available at https://github.com/lunanbit/FedUL.

****

On the Effectiveness of Quasi Character-Level Models for Machine Translation

Salvador Carrión Ponz,Francisco Casacuberta Nolla

Neural Machine Translation (NMT) models often use subword-level vocabularies to deal with rare or unknown words. Although some studies have shown the effectiveness of purely character-based models, these approaches have resulted in highly expensive models in computational terms. In this work, we explore the advantages of quasi character-level Transformers for low-resource NMT, as well as their abi

lity to mitigate the catastrophic forgetting problem. We first present an empirical study on the effectiveness of these models as a function of the size of the training set. As a result, we found that for data-poor environments, quasi character-level Transformers present a competitive advantage over their large subword-level versions. Similarly, we study the generalization of this phenomenon in different languages, domains, and neural architectures. Finally, we conclude this work by studying the ability of these models to mitigate the effects of catastrophic forgetting in machine translation. Our work suggests that quasi character-level Transformers have a competitive advantage in data-poor environments and, although they do not mitigate the catastrophic forgetting problem, they greatly help to achieve greater consistency between domains.
****

Contact Points Discovery for Soft-Body Manipulations with Differentiable Physics
Sizhe Li,Zhiao Huang,Tao Du,Hao Su,Joshua B. Tenenbaum,Chuang Gan

Differentiable physics has recently been shown as a powerful tool for solving soft-body manipulation tasks. However, the differentiable physics solver often gets stuck when the initial contact points of the end effectors are sub-optimal or when performing multi-stage tasks that require contact point switching, which often leads to local minima.

To address this challenge, we propose a contact point discovery approach (CPDeform) that guides the stand-alone differentiable physics solver to deform various soft-body plasticines. The key idea of our approach is to integrate optimal transport-based contact points discovery into the differentiable physics solver to overcome the local minima from initial contact points or contact switching.

On single-stage tasks, our method can automatically find suitable initial contact points based on transport priorities. On complex multi-stage tasks, we can iteratively switch the contact points of end-effectors based on transport priorities. To evaluate the effectiveness of our method, we introduce PlasticineLab-M that extends the existing differentiable physics benchmark PlasticineLab to seven new challenging multi-stage soft-body manipulation tasks. Extensive experimental results suggest that: 1) on multi-stage tasks that are infeasible for the vanilla differentiable physics solver, our approach discovers contact points that efficiently guide the solver to completion; 2) on tasks where the vanilla solver performs sub-optimally or near-optimally, our contact point discovery method performs better than or on par with the manipulation performance obtained with handcrafted contact points.

****

A Study of Aggregation of Long Time-series Input for LSTM Neural Networks
Nitzan Farhi,Yuval Shavitt

Time series forecasting is the process of using time series data to create a prediction model.

Long-short term memory (LSTM) models are the state-of-the-art for time-series forecasting.

However, LSTMs can handle limited length input mostly since when the samples enter the model in sequence,

the oldest samples need to propagate through the LSTM cells self loop for each new sample and thus their data diminishes in this process.

This limits the length of the history that can be used in the training for each time epoch. The common way of handling this problem is by partitioning time records to uniform intervals, averaging each interval, and feeding the LSTM with rather short sequences, but each represents data from a longer history.

In this paper, we show that this common data aggregation method is far from optimal. We generalize the method of partitioning the data, and suggest an Exponential partitioning. We show that non-uniformly partitioning, and especially Exponential partitioning improves LSTM accuracy, significantly. Using other aggregation functions (such as median or maximum) are shown to further improve the accuracy. Overall, using 7 public datasets we show an improvement in accuracy by 6% to 2

7%.
****

Leveraging Automated Unit Tests for Unsupervised Code Translation
Baptiste Roziere,Jie Zhang,Francois Charton,Mark Harman,Gabriel Synnaeve,Guillaume Lample
With little to no parallel data available for programming languages, unsupervised methods are well-suited to source code translation. However, the majority of unsupervised machine translation approaches rely on back-translation, a method developed in the context of natural language translation and one that inherently involves training on noisy inputs. Unfortunately, source code is highly sensitive to small changes; a single token can result in compilation failures or erroneous programs, unlike natural languages where small inaccuracies may not change the meaning of a sentence. To address this issue, we propose to leverage an automated unit-testing system to filter out invalid translations, thereby creating a fully tested parallel corpus. We found that fine-tuning an unsupervised model with this filtered data set significantly reduces the noise in the translations so-generated, comfortably outperforming the state-of-the-art for all language pairs studied. In particular, for Java→Python and Python→C++ we outperform the best previous methods by more than 16% and 24% respectively, reducing the error rate by more than 35%.
****

Transformer Embeddings of Irregularly Spaced Events and Their Participants
Hongyuan Mei,Chenghao Yang,Jason Eisner
The neural Hawkes process (Mei & Eisner, 2017) is a generative model of irregularly spaced sequences of discrete events. To handle complex domains with many event types, Mei et al. (2020a) further consider a setting in which each event in the sequence updates a deductive database of facts (via domain-specific pattern-matching rules); future events are then conditioned on the database contents. They show how to convert such a symbolic system into a neuro-symbolic continuous-time generative model, in which each database fact and possible event has a time-varying embedding that is derived from its symbolic provenance.

In this paper, we modify both models, replacing their recurrent LSTM-based architectures with flatter attention-based architectures (Vaswani et al., 2017), which are simpler and more parallelizable. This does not appear to hurt our accuracy, which is comparable to or better than that of the original models as well as (where applicable) previous attention-based methods (Zuo et al., 2020; Zhang et al., 2020a).
****

Fast Model Editing at Scale
Eric Mitchell,Charles Lin,Antoine Bosselut,Chelsea Finn,Christopher D Manning
While large pre-trained models have enabled impressive results on a variety of downstream tasks, the largest existing models still make errors, and even accurate predictions may become outdated over time. Because detecting all such failures at training time is impossible, enabling both developers and end users of such models to correct inaccurate outputs while leaving the model otherwise intact is desirable. However, the distributed, black-box nature of the representations learned by large neural networks makes producing such targeted edits difficult. If presented with only a single problematic input and new desired output, fine-tuning approaches tend to overfit; other editing algorithms are either computationally infeasible or simply ineffective when applied to very large models. To enable easy post-hoc editing at scale, we propose Model Editor Networks using Gradient Decomposition (MEND), a collection of small auxiliary editing networks that use a single desired input-output pair to make fast, local edits to a pre-trained model's behavior. MEND learns to transform the gradient obtained by standard fine-tuning, using a low-rank decomposition of the gradient to make the parameterization of this transformation tractable. MEND can be trained on a single GPU in less than a day even for 10 billion+ parameter models; once trained MEND enables rapid application of new edits to the pre-trained model. Our experiments with T5, GPT, BERT, and BART models show that MEND is the only approach to model editin

g that effectively edits the behavior of models with more than 10 billion parameters. Code available at https://sites.google.com/view/mend-editing.

****

EfficientPhys: Enabling Simple, Fast, and Accurate Camera-Based Vitals Measurement

Xin Liu,Brian L. Hill,Ziheng Jiang,Shwetak Patel,Daniel McDuff

Camera-based physiological measurement is a growing field with neural models providing state-the-art-performance. Prior research have explored various "end-to-end'' models; however these methods still require several preprocessing steps. These additional operations are often non-trivial to implement making replication and deployment difficult and can even have a higher computational budget than the "core'' network itself. In this paper, we propose two novel and efficient neural models for camera-based physiological measurement called EfficientPhys that remove the need for face detection, segmentation, normalization, color space transformation or any other preprocessing steps. Using an input of raw video frames, our models achieve state-of-the-art accuracy on three public datasets. We show that this is the case whether using a transformer or convolutional backbone. We further evaluate the latency of the proposed networks and show that our most light weight network also achieves a 33\% improvement in efficiency.

****

Why do embedding spaces look as they do?

Xingzhi Guo,Baojian Zhou,Haochen Chen,Sergiy Verstyuk,Steven Skiena

The power of embedding representations is a curious phenomenon.   For embeddings to work effectively as feature representations, there must exist substantial latent structure inherent in the domain to be encoded. Language vocabularies and Wikipedia topics are human-generated structures that reflect how people organize their world, and what they find important. The structure of the resulting embedding spaces reflects the human evolution of language formation and the cultural processes shaping our world.

This paper studies what the observed structure of embeddings can tell us about the natural processes that generate new knowledge or concepts.   We demonstrate that word and graph embeddings trained on standard datasets using several popular algorithms consistently share two distinct properties: (1) a decreasing neighbor frequency concentration with rank, and (2) specific clustering velocities and power-law based community structures.

We then assess a variety of generative models of embedding spaces by these criteria, and conclude that incremental insertion processes based on the Barabási-Albert network generation process best model the observed phenomenon on language and network data.

****

Eigencurve: Optimal Learning Rate Schedule for SGD on Quadratic Objectives with Skewed Hessian Spectrums

Rui Pan,Haishan Ye,Tong Zhang

Learning rate schedulers have been widely adopted in training deep neural networks. Despite their practical importance, there is a discrepancy between its practice and its theoretical analysis. For instance, it is not known what schedules of SGD achieve best convergence, even for simple problems such as optimizing quadratic objectives. In this paper, we propose Eigencurve, the first family of learning rate schedules that can achieve minimax optimal convergence rates (up to a constant) for SGD on quadratic objectives when the eigenvalue distribution of the underlying Hessian matrix is skewed. The condition is quite common in practice. Experimental results show that Eigencurve can significantly outperform step decay in image classification tasks on CIFAR-10, especially when the number of epochs is small. Moreover, the theory inspires two simple learning rate schedulers for practical applications that can approximate eigencurve.

For some problems, the optimal shape of the proposed schedulers resembles that of cosine decay, which sheds light to the success of cosine decay for such situa tions. For other situations, the proposed schedulers are superior to cosine decay.

****

Scalable Sampling for Nonsymmetric Determinantal Point Processes
Insu Han,Mike Gartrell,Jennifer Gillenwater,Elvis Dohmatob,amin karbasi

A determinantal point process (DPP) on a collection of M items is a model, parameterized by a symmetric kernel matrix, that assigns a probability to every subset of those items. Recent work shows that removing the kernel symmetry constraint, yielding nonsymmetric DPPs (NDPPs), can lead to significant predictive performance gains for machine learning applications. However, existing work leaves open the question of scalable NDPP sampling. There is only one known DPP sampling algorithm, based on Cholesky decomposition, that can directly apply to NDPPs as well. Unfortunately, its runtime is cubic in M, and thus does not scale to large item collections. In this work, we first note that this algorithm can be transformed into a linear-time one for kernels with low-rank structure. Furthermore, we develop a scalable sublinear-time rejection sampling algorithm by constructing a novel proposal distribution. Additionally, we show that imposing certain structural constraints on the NDPP kernel enables us to bound the rejection rate in a way that depends only on the kernel rank. In our experiments we compare the speed of all of these samplers for a variety of real-world tasks.

****

An Autoregressive Flow Model for 3D Molecular Geometry Generation from Scratch
Youzhi Luo,Shuiwang Ji

We consider the problem of generating 3D molecular geometries from scratch. While multiple methods have been developed for generating molecular graphs, generating 3D molecular geometries from scratch is largely under-explored. In this work, we propose G-SphereNet, a novel autoregressive flow model for generating 3D molecular geometries. G-SphereNet employs a flexible sequential generation scheme by placing atoms in 3D space step-by-step. Instead of generating 3D coordinates directly, we propose to determine 3D positions of atoms by generating distances, angles and torsion angles, thereby ensuring both invariance and equivariance properties. In addition, we propose to use spherical message passing and attention mechanism for conditional information extraction. Experimental results show that G-SphereNet outperforms previous methods on random molecular geometry generation and targeted molecule discovery tasks. Our code is publicly available as part of the DIG package (https://github.com/divelab/DIG).

****

Optimized Separable Convolution: Yet Another Efficient Convolution Operator
Tao Wei,Yonghong Tian,Yaowei Wang,Yun Liang,Chang Wen Chen

The convolution operation is the most critical component in recent surge of deep learning research. Conventional 2D convolution needs O(C^{2}K^{2}) parameters to represent, where C is the channel size and K is the kernel size. The amount of parameters has become really costly considering that these parameters increased tremendously recently to meet the needs of demanding applications. Among various implementations of the convolution, separable convolution has been proven to be more efficient in reducing the model size. For example, depth separable convolution reduces the complexity to O(C\cdot(C+K^{2})) while spatial separable convolution reduces the complexity to O(C^{2}K). However, these are considered ad hoc designs which cannot ensure that they can in general achieve optimal separation. In this research, we propose a novel and principled operator called optimized separable convolution by optimal design for the internal number of groups and kernel sizes for general separable convolutions can achieve the complexity of O(C^{\frac{3}{2}}K). When the restriction in the number of separated convolutions can be lifted, an even lower complexity at O(C\cdot\log(CK^{2})) can be achieved. Experimental results demonstrate that the proposed optimized separable convolution is able to achieve an improved performance in terms of accuracy-#Params trade-offs over both conventional, depth-wise, and depth/spatial separable convolutions.

****

On Incorporating Inductive Biases into VAEs

Ning Miao,Emile Mathieu,Siddharth N,Yee Whye Teh,Tom Rainforth

We explain why directly changing the prior can be a surprisingly ineffective mechanism for incorporating inductive biases into variational auto-encoders (VAEs), and introduce a simple and effective alternative approach: Intermediary Latent Space VAEs (InteL-VAEs). InteL-VAEs use an intermediary set of latent variables to control the stochasticity of the encoding process, before mapping these in turn to the latent representation using a parametric function that encapsulates our desired inductive bias(es). This allows us to impose properties like sparsity or clustering on learned representations, and incorporate human knowledge into the generative model. Whereas changing the prior only indirectly encourages behavior through regularizing the encoder, InteL-VAEs are able to directly enforce desired characteristics. Moreover, they bypass the computation and encoder design issues caused by non-Gaussian priors, while allowing for additional flexibility through training of the parametric mapping function. We show that these advantages, in turn, lead to both better generative models and better representations being learned.

****

DiffSkill: Skill Abstraction from Differentiable Physics for Deformable Object Manipulations with Tools

Xingyu Lin,Zhiao Huang,Yunzhu Li,Joshua B. Tenenbaum,David Held,Chuang Gan

We consider the problem of sequential robotic manipulation of deformable objects using tools.

Previous works have shown that differentiable physics simulators provide gradients to the environment state and help trajectory optimization to converge orders of magnitude faster than model-free reinforcement learning algorithms for deformable object manipulation. However, such gradient-based trajectory optimization typically requires access to the full simulator states and can only solve short-horizon, single-skill tasks due to local optima. In this work, we propose a novel framework, named DiffSkill, that uses a differentiable physics simulator for skill abstraction to solve long-horizon deformable object manipulation tasks from sensory observations. In particular, we first obtain short-horizon skills using individual tools from a gradient-based optimizer, using the full state information in a differentiable simulator; we then learn a neural skill abstractor from the demonstration trajectories which takes RGBD images as input. Finally, we plan over the skills by finding the intermediate goals and then solve long-horizon tasks. We show the advantages of our method in a new set of sequential deformable object manipulation tasks compared to previous reinforcement learning algorithms and compared to the trajectory optimizer.

****

On the Existence of Universal Lottery Tickets

Rebekka Burkholz,Nilanjana Laha,Rajarshi Mukherjee,Alkis Gotovos

The lottery ticket hypothesis conjectures the existence of sparse subnetworks of large randomly initialized deep neural networks that can be successfully trained in isolation. Recent work has experimentally observed that some of these tickets can be practically reused across a variety of tasks, hinting at some form of universality. We formalize this concept and theoretically prove that not only do such universal tickets exist but they also do not require further training. Our proofs introduce a couple of technical innovations related to pruning for strong lottery tickets, including extensions of subset sum results and a strategy to leverage higher amounts of depth. Our explicit sparse constructions of universal function families might be of independent interest, as they highlight representational benefits induced by univariate convolutional architectures.

****

Pre-training Molecular Graph Representation with 3D Geometry

Shengchao Liu,Hanchen Wang,Weiyang Liu,Joan Lasenby,Hongyu Guo,Jian Tang

Molecular graph representation learning is a fundamental problem in modern drug and material discovery. Molecular graphs are typically modeled by their 2D topological structures, but it has been recently discovered that 3D geometric information plays a more vital role in predicting molecular functionalities. However, the lack of 3D information in real-world scenarios has significantly impeded the learning of geometric graph representation. To cope with this challenge, we propose the Graph Multi-View Pre-training (GraphMVP) framework where self-supervised learning (SSL) is performed by leveraging the correspondence and consistency between 2D topological structures and 3D geometric views. GraphMVP effectively learns a 2D molecular graph encoder that is enhanced by richer and more discriminative 3D geometry. We further provide theoretical insights to justify the effectiveness of GraphMVP. Finally, comprehensive experiments show that GraphMVP can consistently outperform existing graph SSL methods. Code is available on GitHub: https://github.com/chao1224/GraphMVP.

****

Dynamic Graph Representation Learning via Graph Transformer Networks

Weilin Cong,Yanhong Wu,Yuandong Tian,Mengting Gu,Yinglong Xia,Mehrdad Mahdavi,Chun-cheng Jason Chen

Dynamic graph representation learning is an important task with widespread applications. Previous methods on dynamic graph learning are usually sensitive to noisy graph information such as missing or spurious connections, which can yield degenerated performance and generalization. To overcome this challenge, we propose a Transformer-based dynamic graph learning method named Dynamic Graph Transformer (DGT) with spatial-temporal encoding to effectively learn graph topology and capture implicit links. To improve the generalization ability, we introduce two complementary self-supervised pre-training tasks and show that jointly optimizing the two pre-training tasks results in a smaller Bayesian error rate via an information-theoretic analysis. We also propose a temporal-union graph structure and a target-context node sampling strategy for efficient and scalable training. Extensive experiments on real-world datasets illustrate that DGT presents superior performance compared with several state-of-the-art baselines.

****

PER-ETD: A Polynomially Efficient Emphatic Temporal Difference Learning Method

Ziwei Guan,Tengyu Xu,Yingbin Liang

Emphatic temporal difference (ETD) learning (Sutton et al., 2016) is a successful method to conduct the off-policy value function evaluation with function approximation. Although ETD has been shown to converge asymptotically to a desirable value function, it is well-known that ETD often encounters a large variance so that its sample complexity can increase exponentially fast with the number of iterations. In this work, we propose a new ETD method, called PER-ETD (i.e., PEriodically Restarted-ETD), which restarts and updates the follow-on trace only for a finite period for each iteration of the evaluation parameter. Further, PER-ETD features a design of the logarithmical increase of the restart period with the number of iterations, which guarantees the best trade-off between the variance and bias and keeps both vanishing sublinearly. We show that PER-ETD converges to the same desirable fixed point as ETD, but improves the exponential sample complexity of ETD to be polynomials. Our experiments validate the superior performance of PER-ETD and its advantage over ETD.

****

Neural Simulated Annealing

Alvaro Correia,Daniel E. Worrall,Roberto Bondesan

Simulated annealing (SA) is a stochastic global optimisation technique applicable to a wide range of discrete and continuous variable problems. Despite its simplicity, the development of an effective SA optimiser for a given problem hinges on a handful of carefully handpicked components; namely, neighbour proposal distribution and temperature annealing schedule. In this work, we view SA from a reinforcement learning perspective and frame the proposal distribution as a policy, which can be optimised for higher solution quality given a fixed computational budget. We demonstrate that this Neural SA with such a learnt proposal distribution outperforms SA baselines with hand-selected parameters on a number of problems: Rosenbrock's function, the Knapsack problem, the Bin Packing problem, and the Travelling Salesperson problem. We also show that Neural SA scales well to large problems while again outperforming popular off-the-shelf solvers in terms of solution quality and wall clock time.

****

Generating High-Fidelity Privacy-Conscious Synthetic Patient Data for Causal Effect Estimation with Multiple Treatments

Jingpu Shi,Dong Wang,Gino Tesei,Beau Norgeot

A causal effect can be defined as the comparison of outcomes from two or more alternative treatments. Knowing this treatment effect is critically important in healthcare because it makes it possible to identify the best treatment for a person when more than one option exists. In the past decade, there has been exponentially growing interest in the use of observational data collected as a part of routine healthcare practice to determine the effect of a treatment with causal inference models. Validation of these models, however, has been a challenge because the ground truth is unknown: only one treatment-outcome pair for each person can be observed. There have been multiple efforts to fill this void using synthetic data where the ground truth can be generated. However, to date, these datasets have been severely limited in their utility either by being modeled after small non-representative patient populations, being dissimilar to real target populations, or only providing known effects for two cohorts (treated vs control). In this work, we produced a large-scale and realistic synthetic dataset that supports multiple hypertension treatments, by modeling after a nationwide cohort of more than 250,000 hypertension patients' multi-year history of diagnoses, medications, and laboratory values. We designed a data generation process by combining an adapted ADS-GAN model for fictitious patient information generation and a neural network for treatment outcome generation. Wasserstein distance of 0.35 demonstrates that our synthetic data follows a nearly identical joint distribution to the patient cohort used to generate the data. Our dataset provides ground truth effects for about 30 hypertension treatments on blood pressure outcomes. Patient privacy was a primary concern for this study; the ϵ-identifiability metric, which estimates the probability of actual patients being identified, is 0.008%, ensuring that our synthetic data cannot be used to identify any actual patients. Using our dataset, we tested the bias in causal effect estimation of three well-established models: propensity sore stratification, doubly robust approach (DR) with logistic regression, DR with random forest (RF) classification. Interestingly, we found that while the RF DR outperformed the logistic DR as expected, the best performance actually came from propensity score stratification, despite the theoretical strength of statistical properties of the DR family of models. We believe this dataset will facilitate the additional development, evaluation, and comparison of real-world causal models. The approach we used can be readily extended to other types of diseases in the clinical domain, and to datasets in other domains as well.

****

Taming Sparsely Activated Transformer with Stochastic Experts

Simiao Zuo,Xiaodong Liu,Jian Jiao,Young Jin Kim,Hany Hassan,Ruofei Zhang,Jianfeng Gao,Tuo Zhao

Sparsely activated models (SAMs), such as Mixture-of-Experts (MoE), can easily scale to have outrageously large amounts of parameters without significant increase in computational cost. However, SAMs are reported to be parameter inefficient such that larger models do not always lead to better performance. While most on-going research focuses on improving SAMs models by exploring methods of routing inputs to experts, our analysis reveals that such research might not lead to the solution we expect, i.e., the commonly-used routing methods based on gating mechanisms do not work better than randomly routing inputs to experts. In this paper, we propose a new expert-based model, THOR ($\underline{\textbf{T}}$ransformer wit$\underline{\textbf{H}}$ St$\underline{\textbf{O}}$chastic Expe$\underline{\textbf{R}}$ts). Unlike classic expert-based models, such as the Switch Transformer, experts in THOR are randomly activated for each input during training and inference. THOR models are trained using a consistency regularized loss, where experts learn not only from training data but also from other experts as teachers, such that all the experts make consistent predictions. We validate the effectiveness of THOR on machine translation tasks. Results show that THOR models are more parameter efficient in that they significantly outperform the Transformer and MoE models across various settings. For example, in multilingual translation, THOR outperforms the Switch Transformer by 2 BLEU scores, and obtains the same BLEU score as that of a state-of-the-art MoE model that is 18 times larger. Our code is publicly available at: https://github.com/microsoft/Stochastic-Mixture-of-Experts.

****

Fast and Efficient Once-For-All Networks for Diverse Hardware Deployment

Jun Fang,Li Yang,Chengyao Shen,Hamzah Abdel-Aziz,David Thorsley,Joseph Hassoun

Convolutional neural networks are widely used in practical application in many diverse environments. Each different environment requires a different optimized network to maximize accuracy under its unique hardware constraints and latency requirements. To find models for this varied array of potential deployment targets, once-for-all (OFA) was introduced as a way to simultaneously co-train many models at once, while keeping the total training cost constant. However, the total training cost is very high, requiring up to 1200 GPU-hours. Compound OFA (compOFA) decreased the training cost of OFA by 2$\times$ by coupling model dimensions to reduce the search space of possible models by orders of magnitude, while also simplifying the training procedure.

In this work, we continue the effort to reduce the training cost of OFA methods. While both OFA and compOFA use a pre-trained teacher network, we propose an in-place knowledge distillation procedure to train the super-network simultaneously with the sub-networks. Within this in-place distillation framework, we develop an upper-attentive sample technique that reduces the training cost per epoch while maintaining accuracy. Through experiments on ImageNet, we demonstrate that, we can achieve a $2\times$ - $3\times$ ($1.5\times$ - $1.8\times$) reduction in training time compared to the state of the art OFA and compOFA, respectively, without loss of optimality.

****

How to Improve Sample Complexity of SGD over Highly Dependent Data?

Shaocong Ma,Ziyi Chen,Yi Zhou,Kaiyi Ji,Yingbin Liang

Conventional machine learning applications typically assume that data samples are independently and identically distributed (i.i.d.). However, many practical scenarios naturally involve a data-generating process that produces highly dependent data samples, which are known to heavily bias the stochastic optimization process and slow down the convergence of learning. In this paper, we conduct a fundamental study on how to facilitate the convergence of SGD over highly dependent data using different popular update schemes. Specifically, with a ϕ-mixing model that captures both exponential and polynomial decay of the data dependence over time, we show that SGD with periodic data-subsampling achieves an improved sample complexity over the standard SGD in the full spectrum of the ϕ-mixing data dependence. Moreover, we show that by fully utilizing the data, mini-batch SGD can further substantially improve the sample complexity with highly dependent data. Numerical experiments validate our theory.

****

Hierarchical Variational Memory for Few-shot Learning Across Domains

Yingjun Du,Xiantong Zhen,Ling Shao,Cees G. M. Snoek

Neural memory enables fast adaptation to new tasks with just a few training samples. Existing memory models store features only from the single last layer, which does not generalize well in presence of a domain shift between training and test distributions. Rather than relying on a flat memory, we propose a hierarchical alternative that stores features at different semantic levels. We introduce a hierarchical prototype model, where each level of the prototype fetches corresponding information from the hierarchical memory. The model is endowed with the ability to flexibly rely on features at different semantic levels if the domain shift circumstances so demand. We meta-learn the model by a newly derived hierarchical variational inference framework, where hierarchical memory and prototypes are jointly optimized. To explore and exploit the importance of different semantic levels, we further propose to learn the weights associated with the prototype at each level in a data-driven way, which enables the model to adaptively choose the most generalizable features. We conduct thorough ablation studies to demons trate the effectiveness of each component in our model. The new state-of-the-art performance on cross-domain and competitive performance on traditional few-shot classification further substantiates the benefit of hierarchical variational memory.

****

Bayesian Exploration for Lifelong Reinforcement Learning

Haotian Fu,Shangqun Yu,Michael Littman,George Konidaris

A central question in reinforcement learning (RL) is how to leverage prior knowledge to accelerate learning in new tasks. We propose a Bayesian exploration method for lifelong reinforcement learning (BLRL) that aims to learn a Bayesian posterior that distills the common structure shared across different tasks. We further derive a sample complexity analysis of BLRL in the finite MDP setting. To scale our approach, we propose a variational Bayesian Lifelong Learning (VBLRL) algorithm that is based on Bayesian neural networks, can be combined with recent model-based RL methods, and exhibits backward transfer. Experimental results on three challenging domains show that our algorithms adapt to new tasks faster than state-of-the-art lifelong RL methods.

****

Learning Audio-Visual Speech Representation by Masked Multimodal Cluster Prediction

Bowen Shi,Wei-Ning Hsu,Kushal Lakhotia,Abdelrahman Mohamed

Video recordings of speech contain correlated audio and visual information, providing a strong signal for speech representation learning from the speaker’s lip movements and the produced sound. We introduce Audio-Visual Hidden Unit BERT (AV-HuBERT), a self-supervised representation learning framework for audio-visual speech, which masks multi-stream video input and predicts automatically discovered and iteratively refined multimodal hidden units. AV-HuBERT learns powerful audio-visual speech representation benefiting both lip-reading and automatic speech recognition. On the largest public lip-reading benchmark LRS3 (433 hours), AV-HuBERT achieves 32.5% WER with only 30 hours of labeled data, outperforming the former state-of-the-art approach (33.6%) trained with a thousand times more transcribed video data (31K hours) (Makino et al., 2019). The lip-reading WER is further reduced to 26.9% when using all 433 hours of labeled data from LRS3 and combined with self-training. Using our audio-visual representation on the same benchmark for audio-only speech recognition leads to a 40% relative WER reduction over the state-of-the-art performance (1.3% vs 2.3%). Our code and models are available at https://github.com/facebookresearch/av_hubert.

****

Learning Equivariances and Partial Equivariances From Data

David W. Romero,Suhas Lohit

Group equivariant Convolutional Neural Networks (G-CNNs) constrain features to respect the chosen symmetries, and lead to better generalization when these symmetries appear in the data. However, if the chosen symmetries are not present, group equivariant architectures lead to overly constrained models and worse performance. Frequently, the distribution of the 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 partially is better suited to represent the data. Moreover, relevant symmetries may differ for low and high-level features, e.g., edge orientations in a face, and face poses relative to the camera. As a result, the optimal level of equivariance may differ per layer. In this work, we introduce Partial G-CNNs: a family of equivariant networks able to learn partial and full equivariances from data at every layer end-to-end. Partial G-CNNs retain full equivariance whenever beneficial, e.g., for rotated MNIST, but are able to restrict it whenever it becomes harmful, e.g., for 6 / 9 or natural image classification. Partial G-CNNs perform on par with G-CNNs when full equivariance is necessary, and outperform them otherwise. Our method is applicable to discrete groups, continuous groups and combinations thereof.

****

Neural Tangent Kernel Empowered Federated Learning

Kai Yue,Richeng Jin,Ryan Pilgrim,Chau-Wai Wong,Dror Baron,Huaiyu Dai

Federated learning (FL) is a privacy-preserving paradigm where multiple participants jointly solve a machine learning problem without sharing raw data. Unlike traditional distributed learning, a unique characteristic of FL is statistical heterogeneity, namely, data distributions across participants are different from each other. Meanwhile, recent advances in the interpretation of neural networks have seen a wide use of neural tangent kernel (NTK) for convergence and generalization analyses. In this paper, we propose a novel FL paradigm empowered by the NTK framework. The proposed paradigm addresses the challenge of statistical heterogeneity by transmitting update data that are more expressive than those of the traditional FL paradigms. Specifically, sample-wise Jacobian matrices, rather than model weights/gradients, are uploaded by participants. The server then constructs an empirical kernel matrix to update a global model without explicitly performing gradient descent. We further develop a variant with improved communication efficiency and enhanced privacy. Numerical results show that the proposed paradigm can achieve the same accuracy while reducing the number of communication rounds by an order of magnitude compared to federated averaging.

****

An Explanation of In-context Learning as Implicit Bayesian Inference

Sang Michael Xie,Aditi Raghunathan,Percy Liang,Tengyu Ma

Large language models (LMs) such as GPT-3 have the surprising ability to do in-context learning, where the model learns to do a downstream task simply by conditioning on a prompt consisting of input-output examples. The LM learns from these examples without being explicitly pretrained to learn. Thus, it is unclear what enables in-context learning. In this paper, we study how in-context learning can emerge when pretraining documents have long-range coherence. Here, the LM must infer a latent document-level concept to generate coherent next tokens during pretraining. At test time, in-context learning occurs when the LM also infers a shared latent concept between examples in a prompt. We prove when this occurs despite a distribution mismatch between prompts and pretraining data in a setting where the pretraining distribution is a mixture of HMMs. In contrast to messy large-scale datasets used to train LMs capable of in-context learning, we generate a small-scale synthetic dataset (GINC) where Transformers and LSTMs both exhibit in-context learning. Beyond the theory, experiments on GINC exhibit large-scale real-world phenomena including improved in-context performance with model scaling (despite the same pretraining loss), sensitivity to example order, and instances where zero-shot is better than few-shot in-context learning.

****

Differentiable Scaffolding Tree for Molecule Optimization

Tianfan Fu,Wenhao Gao,Cao Xiao,Jacob Yasonik,Connor W. Coley,Jimeng Sun

The structural design of functional molecules, also called molecular optimization, is an essential chemical science and engineering task with important applications, such as drug discovery. Deep generative models and combinatorial optimization methods achieve initial success but still struggle with directly modeling discrete chemical structures and often heavily rely on brute-force enumeration. The challenge comes from the discrete and non-differentiable nature of molecule structures. To address this, we propose differentiable scaffolding tree (DST) that utilizes a learned knowledge network to convert discrete chemical structures to locally differentiable ones. DST enables a gradient-based optimization on a chemical graph structure by back-propagating the derivatives from the target properties through a graph neural network (GNN). Our empirical studies show the gradient-based molecular optimizations are both effective and sample efficient (in terms of oracle calling number). Furthermore, the learned graph parameters can also provide an explanation that helps domain experts understand the model output. The code repository (including processed data, trained model, demonstration, molecules with the highest property) is available at https://github.com/futianfan/DST.

****

Eliminating Sharp Minima from SGD with Truncated Heavy-tailed Noise

Xingyu Wang,Sewoong Oh,Chang-Han Rhee

The empirical success of deep learning is often attributed to SGD’s mysterious a bility to avoid sharp local minima in the loss landscape, as sharp minima are known to lead to poor generalization. Recently, empirical evidence of heavy-tailed gradient noise was reported in many deep learning tasks; and it was shown in (Simsekli et al., 2019a;b) that SGD can escape sharp local minima under the presence of such heavy-tailed gradient noise, providing a partial solution to the mystery. In this work, we analyze a popular variant of SGD where gradients are truncated above a fixed threshold. We show that it achieves a stronger notion of avoiding sharp minima: it can effectively eliminate sharp local minima entirely from its training trajectory. We characterize the dynamics of truncated SGD driven by heavy-tailed noises. First, we show that the truncation threshold and width of the attraction field dictate the order of the first exit time from the associated local minimum. Moreover, when the objective function satisfies appropriate structural conditions, we prove that as the learning rate decreases, the dynamics of the heavy-tailed truncated SGD closely resemble those of a continuous-time Markov chain that never visits any sharp minima. Real data experiments on deep learning confirm our theoretical prediction that heavy-tailed SGD with gradient clipping finds a flatter local minima and achieves better generalization.

****

Federated Learning via Plurality Vote

Kai Yue,Richeng Jin,Chau-Wai Wong,Huaiyu Dai

Federated learning allows collaborative workers to solve a machine learning problem while preserving data privacy. Recent studies have tackled various challenges in federated learning, but the joint optimization of communication overhead, learning reliability, and deployment efficiency is still an open problem. To this end, we propose a new scheme named federated learning via plurality vote (FedVote). In each communication round of FedVote, workers transmit binary or ternary weights to the server with low communication overhead. The model parameters are aggregated via weighted voting to enhance the resilience against Byzantine attacks. When deployed for inference, the model with binary or ternary weights is resource-friendly to edge devices. We show that our proposed method can reduce quantization error and converges faster compared with the methods directly quantizing the model updates.

****

Learning Fast, Learning Slow: A General Continual Learning Method based on Complementary Learning System

Elahe Arani,Fahad Sarfraz,Bahram Zonooz

Humans excel at continually learning from an ever-changing environment whereas it remains a challenge for deep neural networks which exhibit catastrophic forgetting. The complementary learning system (CLS) theory suggests that the interplay between rapid instance-based learning and slow structured learning in the brain is crucial for accumulating and retaining knowledge. Here, we propose CLS-ER, a novel dual memory experience replay (ER) method which maintains short-term and long-term semantic memories that interact with the episodic memory. Our method employs an effective replay mechanism whereby new knowledge is acquired while aligning the decision boundaries with the semantic memories. CLS-ER does not utilize the task boundaries or make any assumption about the distribution of the data which makes it versatile and suited for ``general continual learning''. Our approach achieves state-of-the-art performance on standard benchmarks as well as more realistic general continual learning settings.

****

Amortized Tree Generation for Bottom-up Synthesis Planning and Synthesizable Molecular Design

Wenhao Gao,Rocío Mercado,Connor W. Coley

Molecular design and synthesis planning are two critical steps in the process of molecular discovery that we propose to formulate as a single shared task of conditional synthetic pathway generation. We report an amortized approach to generate synthetic pathways as a Markov decision process conditioned on a target molecular embedding. This approach allows us to conduct synthesis planning in a bottom-up manner and design synthesizable molecules by decoding from optimized conditional codes, demonstrating the potential to solve both problems of design and sy nthesis simultaneously. The approach leverages neural networks to probabilistically model the synthetic trees, one reaction step at a time, according to reactivity rules encoded in a discrete action space of reaction templates. We train these networks on hundreds of thousands of artificial pathways generated from a pool of purchasable compounds and a list of expert-curated templates. We validate our method with (a) the recovery of molecules using conditional generation, (b) the identification of synthesizable structural analogs, and (c) the optimization of molecular structures given oracle functions relevant to bioactivity and drug discovery.

****

Implicit Regularization of Bregman Proximal Point Algorithm and Mirror Descent on Separable Data

Yan Li,Caleb Ju,Ethan Fang,Tuo Zhao

Bregman proximal point algorithm (BPPA), as one of the centerpieces in the optimization toolbox, has been witnessing emerging applications. With a simple and easy-to-implement update rule, the algorithm bears several compelling intuitions for empirical successes, yet rigorous justifications are still largely unexplored. We study the computational properties of BPPA through classification tasks with separable data, and demonstrate provable algorithmic regularization effects associated with BPPA. We show that BPPA attains a non-trivial margin, which closely depends on the condition number of the distance-generating function inducing the Bregman divergence. We further demonstrate that the dependence on the condition number is tight for a class of problems, thus showing the importance of divergence in affecting the quality of the obtained solutions. In addition, we extend our findings to mirror descent (MD), for which we establish similar connections between the margin and Bregman divergence. We demonstrate through a concrete example, and show BPPA/MD converges in direction to the maximal margin solution with respect to the squared Mahalanobis distance. Our theoretical findings are among the first to demonstrate the benign learning properties of BPPA/MD, and also provide strong corroborations
for a careful choice of divergence in the algorithmic design.

****

FedChain: Chained Algorithms for Near-optimal Communication Cost in Federated Learning

Charlie Hou,Kiran Koshy Thekumparampil,Giulia Fanti,Sewoong Oh

Federated learning (FL) aims to minimize the communication complexity of training a model over heterogeneous data distributed across many clients. A common approach is local methods, where clients take multiple optimization steps over local data before communicating with the server (e.g., FedAvg). Local methods can exploit similarity between clients' data. However, in existing analyses, this comes at the cost of slow convergence in terms of the dependence on the number of communication rounds R. On the other hand, global methods, where clients simply return a gradient vector in each round (e.g., SGD), converge faster in terms of R but fail to exploit the similarity between clients even when clients are homogeneous. We propose FedChain, an algorithmic framework that combines the strengths of local methods and global methods to achieve fast convergence in terms of R while leveraging the similarity between clients. Using FedChain, we instantiate algorithms that improve upon previously known rates in the general convex and PL settings, and are near-optimal (via an algorithm-independent lower bound that we show) for problems that satisfy strong convexity. Empirical results support this theoretical gain over existing methods.

****

Restricted Category Removal from Model Representations using Limited Data

Pratik Mazumder,Pravendra Singh,Mohammed Asad Karim

Deep learning models are trained on multiple categories jointly to solve several real-world problems. However, there can be cases where some of the classes may become restricted in the future and need to be excluded after the model has already been trained on them (Class-level Privacy). It can be due to privacy, ethical or legal concerns. A naive solution is to simply train the model from scratch on the complete training data while leaving out the training samples from the re

stricted classes (FDR - full data retraining). But this can be a very time-consuming process. Further, this approach will not work well if we no longer have access to the complete training data and instead only have access to very few training data. The objective of this work is to remove the information about the restricted classes from the network representations of all layers using limited data without affecting the prediction power of the model for the remaining classes. Simply fine-tuning the model on the limited available training data for the remaining classes will not be able to sufficiently remove the restricted class information, and aggressive fine-tuning on the limited data may also lead to overfitting. We propose a novel solution to achieve this objective that is significantly faster ($\sim200\times$ on ImageNet) than the naive solution. Specifically, we propose a novel technique for identifying the model parameters that are mainly relevant to the restricted classes. We also propose a novel technique that uses the limited training data of the restricted classes to remove the restricted class information from these parameters and uses the limited training data of the remaining classes to reuse these parameters for the remaining classes. The model obtained through our approach behaves as if it was never trained on the restricted classes and performs similar to FDR (which needs the complete training data). We also propose several baseline approaches and compare our approach with them in order to demonstrate its efficacy.
****

What Do We Mean by Generalization in Federated Learning?
Honglin Yuan,Warren Richard Morningstar,Lin Ning,Karan Singhal
Federated learning data is drawn from a distribution of distributions: clients are drawn from a meta-distribution, and their data are drawn from local data distributions. Generalization studies in federated learning should separate performance gaps from unseen client data (out-of-sample gap) from performance gaps from unseen client distributions (participation gap). In this work, we propose a framework for disentangling these performance gaps. Using this framework, we observe and explain differences in behavior across natural and synthetic federated datasets, indicating that dataset synthesis strategy can be important for realistic simulations of generalization in federated learning. We propose a semantic synthesis strategy that enables realistic simulation without naturally partitioned data. Informed by our findings, we call out community suggestions for future federated learning works.
****

S3: Supervised Self-supervised Learning under Label Noise
Chen Feng,Georgios Tzimiropoulos,Ioannis Patras
Despite the large progress in supervised learning with Neural Networks, there are significant challenges in obtaining high-quality, large-scale and accurately labeled datasets. In this context, in this paper we address the problem of classification in the presence of noisy labels and more specifically, both close-set and open-set label noise, that is when the true label of a sample may, or may not belong to the set of the given labels. In the heart of our method is a sample selection mechanism that relies on the consistency between the annotated label of a sample and the distribution of the labels in its neighborhood in the feature space, a relabeling mechanism that relies on the confidence of the classifier across subsequent iterations and a training strategy that trains the encoder both with a self-consistency loss and the classifier-encoder with cross-entropy loss on the selected samples alone. Without bells and whistles, such as co-training so as to reduce the self-confirmation bias, our method significantly surpasses previous methods on both CIFAR10/CIFAR100 with artificial noise and real-world noisy datasets such as WebVision and ANIMAL-10N.
****

Robust Generalization of Quadratic Neural Networks via Function Identification
Kan Xu,Hamsa Bastani,Osbert Bastani
A key challenge facing deep learning is that neural networks are often not robust to shifts in the underlying data distribution. We study this problem from the perspective of the statistical concept of parameter identification. Generalization bounds from learning theory often assume that the test distribution is close

to the training distribution. In contrast, if we can identify the ``true'' parameters, then the model generalizes to arbitrary distribution shifts. However, neural networks are typically overparameterized, making parameter identification impossible. We show that for quadratic neural networks, we can identify the function represented by the model even though we cannot identify its parameters. Thus, we can obtain robust generalization bounds even in the overparameterized setting. We leverage this result to obtain new bounds for contextual bandits and transfer learning with quadratic neural networks. Overall, our results suggest that we can improve robustness of neural networks by designing models that can represent the true data generating process. In practice, the true data generating process is often very complex; thus, we study how our framework might connect to neural module networks, which are designed to break down complex tasks into compositions of simpler ones. We prove robust generalization bounds when individual neural modules are identifiable.

****

Frequency-aware SGD for Efficient Embedding Learning with Provable Benefits

Yan Li,Dhruv Choudhary,Xiaohan Wei,Baichuan Yuan,Bhargav Bhushanam,Tuo Zhao,Guanghui Lan

Embedding learning has found widespread applications in recommendation systems and natural language modeling, among other domains. To learn quality embeddings efficiently, adaptive learning rate algorithms have demonstrated superior empirical performance over SGD, largely accredited to their token-dependent learning rate. However, the underlying mechanism for the efficiency of token-dependent learning rate remains underexplored. We show that incorporating frequency information of tokens in the embedding learning problems leads to provably efficient algorithms, and demonstrate that common adaptive algorithms implicitly exploit the frequency information to a large extent. Specifically, we propose (Counter-based) Frequency-aware Stochastic Gradient Descent, which applies a frequency-dependent learning rate for each token, and exhibits provable speed-up compared to SGD when the token distribution is imbalanced. Empirically, we show the proposed algorithms are able to improve or match the performance of adaptive algorithms on benchmark recommendation tasks and a large-scale industrial recommendation system, closing the performance gap between SGD and adaptive algorithms. Our results are the first to show token-dependent learning rate provably improves convergence for non-convex embedding learning problems.

****

SegTime: Precise Time Series Segmentation without Sliding Window

Li Zeng,Baifan Zhou,Mohammad Al-Rifai,Evgeny Kharlamov

Time series are common in a wide range of domains and tasks such as stock market partitioning, sleep stage labelling, and human activity recognition, where segmentation, i.e. splitting time series into segments that correspond to given categories, is often required. A common approach to segmentation is to sub-sample the time series using a sliding window with a certain length and overlapping stride, to create sub-sequences of fixed length, and then classify these sub-sequences into the given categories. This reduces time series segmentation to classification. However, this approach guarantees to find only approximate breakpoints: the precise breakpoints can appear in sub-sequences, and thus the accuracy of segmentation degrades when labels change fast. Also, it ignores possible long-term dependencies between sub-sequences. We propose a neural networks approach SegTime that finds precise breakpoints, obviates sliding windows, handles long-term dependencies, and it is insensitive to the label changing frequency. SegTime does so, thanks to its bi-pass architecture with several structures that can process information in a multi-scale fashion. We extensively evaluated the effectiveness of SegTime with very promising results.

****

Iterative Refinement Graph Neural Network for Antibody Sequence-Structure Co-design

Wengong Jin,Jeremy Wohlwend,Regina Barzilay,Tommi S. Jaakkola

Antibodies are versatile proteins that bind to pathogens like viruses and stimulate the adaptive immune system. The specificity of antibody binding is determine d by complementarity-determining regions (CDRs) at the tips of these Y-shaped proteins. In this paper, we propose a generative model to automatically design the CDRs of antibodies with enhanced binding specificity or neutralization capabilities. Previous generative approaches formulate protein design as a structure-conditioned sequence generation task, assuming the desired 3D structure is given a priori. In contrast, we propose to co-design the sequence and 3D structure of CDRs as graphs. Our model unravels a sequence autoregressively while iteratively refining its predicted global structure. The inferred structure in turn guides subsequent residue choices. For efficiency, we model the conditional dependence between residues inside and outside of a CDR in a coarse-grained manner. Our method achieves superior log-likelihood on the test set and outperforms previous baselines in designing antibodies capable of neutralizing the SARS-CoV-2 virus.

****

High Fidelity Visualization of What Your Self-Supervised Representation Knows About

Florian Bordes,Randall Balestriero,Pascal Vincent

Discovering what is learned by neural networks remains a challenge. In self-supervised learning, classification is the most common task used to evaluate how good a representation is. However, relying only on such downstream task can limit our understanding of how much information is contained in the representation of a given input. In this work, we study how to visualize representations learned with self-supervised models. We investigate a simple gradient descend based method to match a target representation and show the limitations of such techniques. We overcome these limitations by developing a representation-conditioned diffusion model (RCDM) that is able to generate high-quality inputs that share commonalities with a given representation. We further demonstrate how our model's generation quality is on par with state-of-the-art generative models and how the representation conditioning brings new avenues to analyze and improve self-supervised models.

****

Neurosymbolic Deep Generative Models for Sequence Data with Relational Constraints

Halley Young,Maxwell Du,Osbert Bastani

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 at the cost of 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 program synthesis 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.

****

An Equivalence Between Data Poisoning and Byzantine Gradient Attacks

Sadegh Farhadkhani,Rachid Guerraoui,Lê-Nguyên Hoang,Oscar Villemaud

To address the resilience of distributed learning, the ``Byzantine" literature considers a strong threat model where workers can report arbitrary gradients to the parameter server. While this model helped generate several fundamental results, it has however sometimes been considered unrealistic, when the workers are mostly trustworthy machines. In this paper, we show a surprising equivalence between this model and data poisoning, a threat considered much more realistic. More specifically, we prove that any gradient attack can be reduced to data poisoning in a personalized federated learning system that provides PAC guarantees (which we show are both desirable and realistic in various personalized federated learning contexts such as linear regression and classification). Maybe most importantly, we derive a simple and practical attack that may be constructed against classical personalized federated learning models, and we show both theoretically and empirically the effectiveness of this attack.

****

CIC: Contrastive Intrinsic Control for Unsupervised Skill Discovery

Michael Laskin,Hao Liu,Xue Bin Peng,Denis Yarats,Aravind Rajeswaran,Pieter Abbeel

We introduce Contrastive Intrinsic Control (CIC) - an algorithm for unsupervised skill discovery that maximizes the mutual information between skills and state transitions. In contrast to most prior approaches, CIC uses a decomposition of the mutual information that explicitly incentivizes diverse behaviors by maximizing state entropy. We derive a novel lower bound estimate for the mutual information which combines a particle estimator for state entropy to generate diverse behaviors and contrastive learning to distill these behaviors into distinct skills. We evaluate our algorithm on the Unsupervised Reinforcement Learning Benchmark, 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 on prior unsupervised skill discovery methods by 91% and the next-leading overall exploration algorithm by 26% in terms of downstream task performance.

****

Learning Curves for Gaussian Process Regression with Power-Law Priors and Targets

Hui Jin,Pradeep Kr. Banerjee,Guido Montufar

We characterize the power-law asymptotics of learning curves for Gaussian process regression (GPR) under the assumption that the eigenspectrum of the prior and the eigenexpansion coefficients of the target function follow a power law. Under similar assumptions, we leverage the equivalence between GPR and kernel ridge regression (KRR) to show the generalization error of KRR. Infinitely wide neural networks can be related to GPR with respect to the neural network GP kernel and the neural tangent kernel, which in several cases is known to have a power-law spectrum. Hence our methods can be applied to study the generalization error of infinitely wide neural networks. We present toy experiments demonstrating the theory.

****

Gradient-based Counterfactual Explanations using Tractable Probabilistic Models

Xiaoting Shao,Kristian Kersting

Counterfactual examples are an appealing class of post-hoc explanations for machine learning models. Given input x of class y, its counterfactual is a contrastive example x' of another class y'. Current approaches primarily solve this task by a complex optimization: define an objective function based on the loss of the counterfactual outcome y' with hard or soft constraints, then optimize this function as a black-box. This “deep learning” approach, however, is rather slow, sometimes tricky, and may result in unrealistic counterfactual examples. In this work, we propose a novel approach to deal with these problems using only two gradient computations based on tractable probabilistic models. First, we compute an unconstrained counterfactual u of x to induce the counterfactual outcome y'. Then, we adapt u to higher density regions, resulting in x'. Empirical evidence demonstrates the dominant advantages of our approach.

****

KNIFE: Kernelized-Neural Differential Entropy Estimation

Georg Pichler,Pierre Colombo,Malik Boudiaf,Günther Koliander,Pablo Piantanida

Estimation of (differential) entropy and the related mutual information has been pursued with significant efforts by the machine learning community. To address shortcomings in previously proposed estimators for differential entropy, here we introduce KNIFE, a fully parameterized, differentiable kernel-based estimator of differential entropy. The flexibility of our approach also allows us to constr uct KNIFE-based estimators for conditional (on either discrete or continuous variables) differential entropy, as well as mutual information. We empirically validate our method on high-dimensional synthetic data and further apply it to guide the training of neural networks for real-world tasks. Our experiments on a large variety of tasks, including visual domain adaptation, textual fair classification, and textual fine-tuning demonstrate the effectiveness of KNIFE-based estimation.

****

Fast topological clustering with Wasserstein distance

Tananun Songdechakraiwut,Bryan M Krause,Matthew I Banks,Kirill V Nourski,Barry D Van Veen

The topological patterns exhibited by many real-world networks motivate the development of topology-based methods for assessing the similarity of networks. However, extracting topological structure is difficult, especially for large and dense networks whose node degrees range over multiple orders of magnitude. In this paper, we propose a novel and computationally practical topological clustering method that clusters complex networks with intricate topology using principled theory from persistent homology and optimal transport. Such networks are aggregated into clusters through a centroid-based clustering strategy based on both their topological and geometric structure, preserving correspondence between nodes in different networks. The notions of topological proximity and centroid are characterized using a novel and efficient approach to computation of the Wasserstein distance and barycenter for persistence barcodes associated with connected components and cycles. The proposed method is demonstrated to be effective using both simulated networks and measured functional brain networks.

****

Churn Reduction via Distillation

Heinrich Jiang,Harikrishna Narasimhan,Dara Bahri,Andrew Cotter,Afshin Rostamizadeh

In real-world systems, models are frequently updated as more data becomes available, and in addition to achieving high accuracy, the goal is to also maintain a low difference in predictions compared to the base model (i.e. predictive churn). If model retraining results in vastly different behavior, then it could cause negative effects in downstream systems, especially if this churn can be avoided with limited impact on model accuracy. In this paper, we show an equivalence between training with distillation using the base model as the teacher and training with an explicit constraint on the predictive churn. We then show that distillation performs strongly for low churn training against a number of recent baselines on a wide range of datasets and model architectures, including fully-connected networks, convolutional networks, and transformers.

****

Learning Causal Models from Conditional Moment Restrictions by Importance Weighting

Masahiro Kato,Masaaki Imaizumi,Kenichiro McAlinn,Shota Yasui,Haruo Kakehi

We consider learning causal relationships under conditional moment restrictions. Unlike causal inference under unconditional moment restrictions, conditional moment restrictions pose serious challenges for causal inference. To address this issue, we propose a method that transforms conditional moment restrictions to unconditional moment restrictions through importance weighting using a conditional density ratio estimator. Then, using this transformation, we propose a method that successfully estimate a parametric or nonparametric functions defined under the conditional moment restrictions. We analyze the estimation error and provide a bound on the structural function, providing theoretical support for our proposed method. In experiments, we confirm the soundness of our proposed method.

****

Mean-Variance Efficient Reinforcement Learning by Expected Quadratic Utility Maximization

Masahiro Kato,Kei Nakagawa,Kenshi Abe,Tetsuro Morimura

Risk management is critical in decision making, and mean-variance (MV) trade-off is one of the most common criteria. However, in reinforcement learning (RL) for

sequential decision making under uncertainty, most of the existing methods for MV control suffer from computational difficulties owing to calculating the gradient of the variance term. In this paper, in contrast to strict MV control, we consider learning MV efficient policies that achieve Pareto efficiency regarding MV trade-off. To achieve this purpose, we train an agent to maximize the expected quadratic utility function, a common objective of risk management in finance and economics. We call our approach RL based on expected quadratic utility maximization (EQUMRL). The EQUMRL does not suffer from the computational difficulties because it does not include gradient estimation of the variance. We confirm that the maximizer of the objective in the EQUMRL directly corresponds to an MV efficient policy under a certain condition. We conduct experiments with benchmark settings to demonstrate the effectiveness of the EQUMRL.
****

Better state exploration using action sequence equivalence
Nathan Grinsztajn,Toby Johnstone,Johan Ferret,Philippe Preux
Incorporating prior knowledge in reinforcement learning algorithms is mainly an open question. Even when insights about the environment dynamics are available, reinforcement learning is traditionally used in a \emph{tabula rasa} setting and must explore and learn everything from scratch. In this paper, we consider the problem of exploiting priors about action sequence equivalence: that is, when different sequences of actions produce the same effect. We propose a new local exploration strategy calibrated to minimize collisions and maximize new state visitations. We show that this strategy can be computed at little cost, by solving a convex optimization problem. By replacing the usual ϵ-greedy strategy in a DQN, we demonstrate its potential in several environments with various dynamic structures.
****

Autonomous Reinforcement Learning: Formalism and Benchmarking
Archit Sharma,Kelvin Xu,Nikhil Sardana,Abhishek Gupta,Karol Hausman,Sergey Levine,Chelsea Finn
Reinforcement learning (RL) provides a naturalistic framing for learning through trial and error, which is appealing both because of its simplicity and effectiveness and because of its resemblance to how humans and animals acquire skills through experience. However, real-world embodied learning, such as that performed by humans and animals, is situated in a continual, non-episodic world, whereas common benchmark tasks in RL are episodic, with the environment resetting between trials to provide the agent with multiple attempts. This discrepancy presents a major challenge when we attempt to take RL algorithms developed for episodic simulated environments and run them on real-world platforms, such as robots. In this paper, we aim to address this discrepancy by laying out a framework for Autonomous Reinforcement Learning (ARL): reinforcement learning where the agent not only learns through its own experience, but also contends with lack of human supervision to reset between trials. We introduce a simulated benchmark EARL based on this framework, containing a set of diverse and challenging simulated tasks reflective of the hurdles introduced to learning when only a minimal reliance on extrinsic intervention can be assumed. We show that standard approaches to episodic RL and existing approaches struggle as interventions are minimized, underscoring the need for developing new algorithms for reinforcement learning with a greater focus on autonomy.
****

Local Permutation Equivariance For Graph Neural Networks
Joshua Mitton,Roderick Murray-Smith
In this work we develop a new method, named {\it locally permutation-equivariant graph neural networks}, which provides a framework for building graph neural networks that operate on local node neighbourhoods, through sub-graphs, while using permutation equivariant update functions. The potential benefits of learning on graph-structured data are vast, and relevant to many application domains. However, one of the challenges, is that graphs are not always of the same size, and often each node in a graph does not have the same connectivity. This necessitates that the update function must be flexible to the input size, which is not the

case in most other domains.

Using our locally permutation-equivariant graph neural networks ensures an expressive update function through using permutation representations, while operating on a lower-dimensional space than that utilised in global permutation equivariance. Furthermore, the use of local update functions offers a significant improvement in GPU memory over global methods. We demonstrate that our method can outperform competing methods on a set of widely used graph benchmark classification tasks.

****

GRAND++: Graph Neural Diffusion with A Source Term

Matthew Thorpe,Tan Minh Nguyen,Hedi Xia,Thomas Strohmer,Andrea Bertozzi,Stanley Osher,Bao Wang

We propose GRAph Neural Diffusion with a source term (GRAND++) for graph deep learning with a limited number of labeled nodes, i.e., low-labeling rate. GRAND++ is a class of continuous-depth graph deep learning architectures whose theoretical underpinning is the diffusion process on graphs with a source term. The source term guarantees two interesting theoretical properties of GRAND++: (i) the representation of graph nodes, under the dynamics of GRAND++, will not converge to a constant vector over all nodes even as the time goes to infinity, which mitigates the over-smoothing issue of graph neural networks and enables graph learning in very deep architectures. (ii) GRAND++ can provide accurate classification even when the model is trained with a very limited number of labeled training data. We experimentally verify the above two advantages on various graph deep learning benchmark tasks, showing a significant improvement over many existing graph neural networks.

****

Case-based reasoning for better generalization in textual reinforcement learning

Mattia Atzeni,Shehzaad Zuzar Dhuliawala,Keerthiram Murugesan,Mrinmaya Sachan

Text-based games (TBG) have emerged as promising environments for driving research in grounded language understanding and studying problems like generalization and sample efficiency. Several deep reinforcement learning (RL) methods with varying architectures and learning schemes have been proposed for TBGs. However, these methods fail to generalize efficiently, especially under distributional shifts. In a departure from deep RL approaches, in this paper, we propose a general method inspired by case-based reasoning to train agents and generalize out of the training distribution. The case-based reasoner collects instances of positive experiences from the agent's interaction with the world and later reuses the collected experiences to act efficiently. The method can be used in conjunction with any existing on-policy neural agent introduced in the literature for TBGs. Our experiments show that the proposed approach consistently improves existing methods, obtains good out-of-distribution generalization and achieves new state-of-the-art results on widely used environments.

****

Constrained Discrete Black-Box Optimization using Mixed-Integer Programming

Theodore Papalexopoulos,Christian Tjandraatmadja,Ross Anderson,Juan Pablo Vielma,David Benjamin Belanger

Discrete black-box optimization problems are challenging for model-based optimization (MBO) algorithms, such as Bayesian optimization, due to the size of the search space and the need to satisfy combinatorial constraints. In particular, these methods require repeatedly solving a complex discrete global optimization problem in the inner loop, where popular heuristic inner-loop solvers introduce approximations and are difficult to adapt to combinatorial constraints. In response, we propose NN+MILP, a general discrete MBO framework using piecewise-linear neural networks as surrogate models and mixed-integer linear programming (MILP) to optimize the acquisition function. MILP provides optimality guarantees and a versatile declarative language for domain-specific constraints. We test our approach on a range of unconstrained and constrained problems, including DNA binding and the NAS-Bench-101 neural architecture search benchmark. NN+MILP surpasses or matches the performance of algorithms tailored to the domain at hand, with glob al optimization of the acquisition problem running in a few minutes using only standard software packages and hardware.

****

Beyond Faithfulness: A Framework to Characterize and Compare Saliency Methods

Angie Boggust,Harini Suresh,Hendrik Strobelt,John Guttag,Arvind Satyanarayan

Saliency methods calculate how important each input feature is to a machine learning model’s prediction, and are commonly used to understand model reasoning. “Faithfulness,” or how fully and accurately the saliency output reflects the underlying model, is an oft-cited desideratum for these methods. However, explanation methods must necessarily sacrifice certain information in service of user-oriented goals such as simplicity. To that end, and akin to performance metrics, we frame saliency methods as abstractions: individual tools that provide insight into specific aspects of model behavior and entail tradeoffs. Using this framing, we describe a framework of nine dimensions to characterize and compare the properties of saliency methods. We group these dimensions into three categories that map to different phases of the interpretation process: methodology, or how the saliency is calculated; sensitivity, or relationships between the saliency result and the underlying model or input; and, perceptibility, or how a user interprets the result. As we show, these dimensions give us a granular vocabulary for describing and comparing saliency methods — for instance, allowing us to develop “saliency cards” as a form of documentation, or helping downstream users understand tradeoffs and choose a method for a particular use case. Moreover, by situating existing saliency methods within this framework, we identify opportunities for future work, including filling gaps in the landscape and developing new evaluation metrics.

****

Neural Deep Equilibrium Solvers

Shaojie Bai,Vladlen Koltun,J Zico Kolter

A deep equilibrium (DEQ) model abandons traditional depth by solving for the fixed point of a single nonlinear layer f_θ. This structure enables decoupling the internal structure of the layer (which controls representational capacity) from how the fixed point is actually computed (which impacts inference-time efficiency), which is usually via classic techniques such as Broyden's method or Anderson acceleration. In this paper, we show that one can exploit such decoupling and substantially enhance this fixed point computation using a custom neural solver. Specifically, our solver uses a parameterized network to both guess an initial value of the optimization and perform iterative updates, in a method that generalizes a learnable form of Anderson acceleration and can be trained end-to-end in an unsupervised manner. Such a solution is particularly well suited to the implicit model setting, because inference in these models requires repeatedly solving for a fixed point of the same nonlinear layer for different inputs, a task at which our network excels. Our experiments show that these neural equilibrium solvers are fast to train (only taking an extra 0.9-1.1% over the original DEQ's training time), require few additional parameters (1-3% of the original model size), yet lead to a $2\times$ speedup in DEQ network inference without any degradation in accuracy across numerous domains and tasks.

****

Variable Length Variable Quality Audio Steganography

Seungmo Ku

Steganography is the task of hiding and recovering secret data inside a non-secret container data while making imperceptible changes to the container. When using steganography to hide audio inside an image, current approaches neither allow the encoding of a signal with variable length nor allow making a trade-off between secret data reconstruction quality and imperceptibility in the changes made to the container image. To address this problem, we propose VLVQ (Variable Length Variable Quality Audio Steganography), a deep learning based steganographic framework capable of hiding variable-length audio inside an image by training the network to iteratively encode and decode the audio data from the container image. Complementary to the standard reconstruction loss, we propose an optional condi tional loss term that allows the users to make quality trade-offs between audio and image reconstruction on inference time, without needing to train a separate model for each trade-off setups. Our experiments on ImageNet and AudioSet demonstrate VLVQ’s ability to retain reasonable image quality (28.99 $psnr$) and audio reconstruction quality (23.79 $snrseg$) while encoding 19 seconds of audio. We also show VLVQ’s capability to generalize to signals longer than what is seen during training.

****

COLA: Consistent Learning with Opponent-Learning Awareness

Timon Willi,Johannes Treutlein,Alistair Letcher,Jakob Nicolaus Foerster

Optimization problems with multiple, interdependent losses, such as Generative Adversarial Networks (GANs) or multi-agent RL, are commonly formalized as differentiable games.

Learning with Opponent-Learning Awareness (LOLA) introduced opponent shaping to this setting. More specifically, LOLA introduced an augmented learning rule that accounts for the agent's influence on the anticipated learning step of the other agents. However, the original LOLA formulation is inconsistent because LOLA models other agents as naive learners rather than LOLA agents.

In previous work, this inconsistency was stated to be the root cause of LOLA's failure to preserve stable fixed points (SFPs). We provide a counterexample by investigating cases where Higher-Order LOLA (HOLA) converges.

Furthermore, we show that, contrary to claims made, Competitive Gradient Descent (CGD) does not solve the consistency problem.

Next, we propose a new method called Consistent LOLA (COLA), which learns update functions that are consistent under mutual opponent shaping. Lastly, we empirically compare the performance and consistency of HOLA, LOLA, and COLA on a set of general-sum learning games.

****

Physics Informed Convex Artificial Neural Networks (PICANNs) for Optimal Transport based Density Estimation

Amanpreet Singh,Martin Bauer,Sarang Joshi

Optimal Mass Transport (OMT) is a well studied problem with a variety of applications in a diverse set of fields ranging from Physics to Computer Vision and in particular Statistics and Data Science. Since the original formulation of Monge in 1781 significant theoretical progress been made on the existence, uniqueness and properties of the optimal transport maps. The actual numerical computation of the transport maps, particularly in high dimensions, remains a challenging problem. In the past decade several neural network based algorithms have been proposed to tackle this task. In this paper, building on recent developments of input convex neural networks and physics informed neural networks for solving PDE's, we propose a new Deep Learning approach to solve the continuous OMT problem. Our framework is based on Brenier's theorem, which reduces the continuous OMT problem to that of solving a non-linear PDE of Monge-Ampere type whose solution is a convex function. To demonstrate the accuracy of our framework we compare our method to several other deep learning based algorithms. We then focus on applications to the ubiquitous density estimation and generative modeling tasks in statistics and machine learning. Finally as an example we present how our framework can be incorporated with an autoencoder to estimate an effective probabilistic generative model.

****

Statistically Meaningful Approximation: a Theoretical Analysis for Approximating Turing Machines with Transformers

Colin Wei,Yining Chen,Tengyu Ma

A common lens to theoretically study neural net architectures is to analyze the functions they can approximate. However, constructions from approximation theory may be unrealistic and therefore less meaningful. For example, a common unrealistic trick is to encode target function values using infinite precision. To address these issues, this work proposes a formal definition of statistically meaningful (SM) approximation which requires the approximating network to exhibit good statistical learnability. We study SM approximation for two function classes: b oolean circuits and Turing machines. We show that overparameterized feedforward neural nets can SM approximate boolean circuits with sample complexity depending only polynomially on the circuit size, not the size of the network. In addition, we show that transformers can SM approximate Turing machines with computation time bounded by T with sample complexity polynomial in the alphabet size, state space size, and $log(T)$. We also introduce new tools for analyzing generalization which provide much tighter sample complexities than the typical VC-dimension or norm-based bounds, which may be of independent interest.

****

Classification and Uncertainty Quantification of Corrupted Data using Semi-Supervised Autoencoders

Philipp Joppich,Sebastian Dorn,Oliver De Candido,Wolfgang Utschick,Jakob Knollmüller

Parametric and non-parametric classifiers often have to deal with real-world data, where corruptions like noise, occlusions, and blur are unavoidable – posing significant challenges. We present a probabilistic approach to classify strongly corrupted data and quantify uncertainty, despite the model only having been trained with uncorrupted data. A semi-supervised autoencoder trained on uncorrupted data is the underlying architecture. We use the decoding part as a generative model for realistic data and extend it by convolutions, masking, and additive Gaussian noise to describe imperfections. This constitutes a statistical inference task in terms of the optimal latent space activations of the underlying uncorrupted datum. We solve this problem approximately with Metric Gaussian Variational Inference (MGVI). The supervision of the autoencoder’s latent space allows us to classify corrupted data directly under uncertainty with the statistically inferred latent space activations. Furthermore, we demonstrate that the model uncertainty strongly depends on whether the classification is correct or wrong, setting a basis for a statistical "lie detector" of the classification. Independent of that, we show that the generative model can optimally restore the uncorrupted datum by decoding the inferred latent space activations.

****

A Theoretical Analysis on Feature Learning in Neural Networks: Emergence from Inputs and Advantage over Fixed Features

Zhenmei Shi,Junyi Wei,Yingyu Liang

An important characteristic of neural networks is their ability to learn representations of the input data with effective features for prediction, which is believed to be a key factor to their superior empirical performance. To better understand the source and benefit of feature learning in neural networks, we consider learning problems motivated by practical data, where the labels are determined by a set of class relevant patterns and the inputs are generated from these along with some background patterns. We prove that neural networks trained by gradient descent can succeed on these problems. The success relies on the emergence and improvement of effective features, which are learned among exponentially many candidates efficiently by exploiting the data (in particular, the structure of the input distribution). In contrast, no linear models on data-independent features of polynomial sizes can learn to as good errors. Furthermore, if the specific input structure is removed, then no polynomial algorithm in the Statistical Query model can learn even weakly. These results provide theoretical evidence showing that feature learning in neural networks depends strongly on the input structure and leads to the superior performance. Our preliminary experimental results on synthetic and real data also provide positive support.

****

Dynamic and Efficient Gray-Box Hyperparameter Optimization for Deep Learning

Martin Wistuba,Arlind Kadra,Josif Grabocka

Gray-box hyperparameter optimization techniques have recently emerged as a promising direction for tuning Deep Learning methods. However, the multi-budget search mechanisms of existing prior works can suffer from the poor correlation among the performances of hyperparameter configurations at different budgets. As a remedy, we introduce DyHPO, a method that learns to dynamically decide which configuration to try next, and for what budget. Our technique is a modification to the classical Bayesian optimization for a gray-box setup. Concretely, we propose a new surrogate for Gaussian Processes that embeds the learning curve dynamics and a new acquisition function that incorporates multi-budget information. We demonstrate the significant superiority of DyHPO against state-of-the-art hyperparameter optimization baselines through large-scale experiments comprising 50 datasets (Tabular, Image, NLP) and diverse neural networks (MLP, CNN/NAS, RNN).

****

CADDA: Class-wise Automatic Differentiable Data Augmentation for EEG Signals

Cédric Rommel,Thomas Moreau,Joseph Paillard,Alexandre Gramfort

Data augmentation is a key element of deep learning pipelines, as it informs the network during training about transformations of the input data that keep the label unchanged. Manually finding adequate augmentation methods and parameters for a given pipeline is however rapidly cumbersome. In particular, while intuition can guide this decision for images, the design and choice of augmentation policies remains unclear for more complex types of data, such as neuroscience signals. Besides, class-dependent augmentation strategies have been surprisingly unexplored in the literature, although it is quite intuitive: changing the color of a car image does not change the object class to be predicted, but doing the same to the picture of an orange does. This paper investigates gradient-based automatic data augmentation algorithms amenable to class-wise policies with exponentially larger search spaces. Motivated by supervised learning applications using EEG signals for which good augmentation policies are mostly unknown, we propose a new differentiable relaxation of the problem. In the class-agnostic setting, results show that our new relaxation leads to optimal performance with faster training than competing gradient-based methods, while also outperforming gradient-free methods in the class-wise setting. This work proposes also novel differentiable augmentation operations relevant for sleep stage classification.

****

Label Leakage and Protection in Two-party Split Learning

Oscar Li,Jiankai Sun,Xin Yang,Weihao Gao,Hongyi Zhang,Junyuan Xie,Virginia Smith,Chong Wang

Two-party split learning is a popular technique for learning a model across feature-partitioned data. In this work, we explore whether it is possible for one party to steal the private label information from the other party during split training, and whether there are methods that can protect against such attacks. Specifically, we first formulate a realistic threat model and propose a privacy loss metric to quantify label leakage in split learning. We then show that there exist two simple yet effective methods within the threat model that can allow one party to accurately recover private ground-truth labels owned by the other party. To combat these attacks, we propose several random perturbation techniques, including $\texttt{Marvell}$, an approach that strategically finds the structure of the noise perturbation by minimizing the amount of label leakage (measured through our quantification metric) of a worst-case adversary. We empirically demonstrate the effectiveness of our protection techniques against the identified attacks, and show that $\texttt{Marvell}$ in particular has improved privacy-utility tradeoffs relative to baseline approaches.

****

Tight lower bounds for Differentially Private ERM

Daogao Liu,Zhou Lu

We consider the lower bounds of differentially private ERM for general convex functions. For approximate-DP, the well-known upper bound of DP-ERM is $O(\frac{\sqrt{p\log(1/\delta)}}{\epsilon n})$, which is believed to be tight. However, current lower bounds are off by some logarithmic terms, in particular $\Omega(\frac{\sqrt{p}}{\epsilon n})$ for constrained case and $\Omega(\frac{\sqrt{p}}{\epsilon n \log p})$ for unconstrained case.

We achieve tight $\Omega(\frac{\sqrt{p \log(1/\delta)}}{\epsilon n})$ lower bounds for both cases by introducing a novel biased mean property for fingerprinting codes. As for pure-DP, we utilize a novel ℓ_2 loss function instead of linear functions considered by previous papers, and achieve the first (tight) $\Ome ga(\frac{p}{\epsilon n})$ lower bound. We also introduce an auxiliary dimension to simplify the computation brought by ℓ_2 loss. Our results close a gap in our understanding of DP-ERM by presenting the fundamental limits. Our techniques may be of independent interest, which help enrich the tools so that it readily applies to problems that are not (easily) reducible from one-way marginals.

****

Semi-relaxed Gromov-Wasserstein divergence and applications on graphs
Cédric Vincent-Cuaz,Rémi Flamary,Marco Corneli,Titouan Vayer,Nicolas Courty

Comparing structured objects such as graphs is a fundamental operation involved in many learning tasks. To this end, the Gromov-Wasserstein (GW) distance, based on Optimal Transport (OT), has proven to be successful in handling the specific nature of the associated objects. More specifically, through the nodes connectivity relations, GW operates on graphs, seen as probability measures over specific spaces. At the core of OT is the idea of conservation of mass, which imposes a coupling between all the nodes from the two considered graphs. We argue in this paper that this property can be detrimental for tasks such as graph dictionary or partition learning, and we relax it by proposing a new semi-relaxed Gromov-Wasserstein divergence. Aside from immediate computational benefits, we discuss its properties, and show that it can lead to an efficient graph dictionary learning algorithm. We empirically demonstrate its relevance for complex tasks on graphs such as partitioning, clustering and completion.

****

CodeTrek: Flexible Modeling of Code using an Extensible Relational Representation
Pardis Pashakhanloo,Aaditya Naik,Yuepeng Wang,Hanjun Dai,Petros Maniatis,Mayur Naik

Designing a suitable representation for code-reasoning tasks is challenging in aspects such as the kinds of program information to model, how to combine them, and how much context to consider. We propose CodeTrek, a deep learning approach that addresses these challenges by representing codebases as databases that conform to rich relational schemas. The relational representation not only allows CodeTrek to uniformly represent diverse kinds of program information, but also to leverage program-analysis queries to derive new semantic relations, which can be readily incorporated without further architectural engineering. CodeTrek embeds this relational representation using a set of walks that can traverse different relations in an unconstrained fashion, and incorporates all relevant attributes along the way. We evaluate CodeTrek on four diverse and challenging Python tasks: variable misuse, exception prediction, unused definition, and variable shadowing. CodeTrek achieves an accuracy of 91%, 63%, 98%, and 94% on these tasks respectively, and outperforms state-of-the-art neural models by 2-19% points.

****

Bridging Recommendation and Marketing via Recurrent Intensity Modeling
Yifei Ma,Ge Liu,Anoop Deoras

This paper studies some under-explored connections between personalized recommendation and marketing systems. Obviously, these two systems are different, in two main ways. Firstly, personalized item-recommendation (ItemRec) is user-centric, whereas marketing recommends the best user-state segments (UserRec) on behalf of its item providers. (We treat different temporal states of the same user as separate marketing opportunities.) To overcome this difference, we realize a novel connection to Marked-Temporal Point Processes (MTPPs), where we view both problems as different projections from a unified temporal intensity model for all user-item pairs. Correspondingly, we derive Recurrent Intensity Models (RIMs) to extend from recurrent ItemRec models with minimal changes. The second difference between recommendation and marketing is in the temporal domains where they operate. While recommendation demands immediate responses in real-time, marketing campaigns are often long-term, setting goals to cover a given percentage of all opportunities for a given item in a given period of time. We formulate both considerations into a constrained optimization problem we call online match (OnlnMtch) and derive a solution we call Dual algorithm. Simply put, Dual modifies the real-time ItemRec scores such that the marketing constraints can be met with least compromises in user-centric utilities. Finally, our connections between recommendation and marketing may lead to novel applications. We run experiments where we use marketing as an alternative to cold-start item exploration, by setting a minimal-exposure constraint for every item in the audience base. Our experiments are available at \url{https://github.com/awslabs/recurrent-intensity-model-experiments}

****

A Fine-Grained Analysis on Distribution Shift

Olivia Wiles,Sven Gowal,Florian Stimberg,Sylvestre-Alvise Rebuffi,Ira Ktena,Krishnamurthy Dj Dvijotham,Ali Taylan Cemgil

Robustness to distribution shifts is critical for deploying machine learning models in the real world. Despite this necessity, there has been little work in defining the underlying mechanisms that cause these shifts and evaluating the robustness of algorithms across multiple, different distribution shifts. To this end, we introduce a framework that enables fine-grained analysis of various distribution shifts. We provide a holistic analysis of current state-of-the-art methods by evaluating 19 distinct methods grouped into five categories across both synthetic and real-world datasets. Overall, we train more than 85K models. Our experimental framework can be easily extended to include new methods, shifts, and datasets. We find, unlike previous work (Gulrajani & Lopez-Paz, 2021), that progress has been made over a standard ERM baseline; in particular, pretraining and augmentations (learned or heuristic) offer large gains in many cases. However, the best methods are not consistent over different datasets and shifts. We will open source our experimental framework, allowing future work to evaluate new methods over multiple shifts to obtain a more complete picture of a method's effectiveness.

Code is available at github.com/deepmind/distribution_shift_framework.

****

Sparse Attention with Learning to Hash

Zhiqing Sun,Yiming Yang,Shinjae Yoo

Transformer has become ubiquitous in sequence modeling tasks. As a key component of Transformer, self-attention does not scale to long sequences due to its quadratic time and space complexity with respect to the sequence length. To tackle this problem, recent work developed dynamic attention sparsification techniques based on Approximate Nearest Neighbor (ANN) methods, where similar queries and keys are allocated to the same hash bucket with high probability. However, the effectiveness of those ANN methods relies on the assumption that queries and keys should lie in the same space, which is not well justified. Besides, some of the ANN methods such as Locality-Sensitive Hashing (LSH) are randomized and cannot fully utilize the available real data distributions. To overcome these issues, this paper proposes a new strategy for sparse attention, namely LHA (Learning-to-Hash Attention), which directly learns separate parameterized hash functions for queries and keys, respectively. Another advantage of LHA is that it does not impose extra constraints for queries and keys, which makes it applicable to the wide range of pre-trained Transformer models. Our experiments on evaluation of the WikiText-103 dataset for language modeling, the GLUE benchmark for natural language understanding, and the Lang-Range-Arena benchmark for multiple tasks (text/image classification, retrieval, etc.) show the superior performance of LHA over other strong Transformer variants.

****

Certified Adversarial Robustness Under the Bounded Support Set

Yiwen Kou,Qinyuan Zheng,Yisen Wang

Deep neural networks (DNNs) have revealed severe vulnerability to adversarial perturbations, beside empirical adversarial training for robustness, the design of provably robust classifiers attracts more and more attention. Randomized smoothing method provides the certified robustness with agnostic architecture, which is further extended to a provable robustness framework using f-divergence. While these methods cannot be applied to smoothing measures with bounded support set such as uniform probability measure due to the use of likelihood ratio in their certification methods. In this paper, we introduce a framework that is able to deal with robustness properties of arbitrary smoothing measures including those with bounded support set by using Wasserstein distance as well as total variation distance. By applying our methodology to uniform probability measures with support set $B_{2}(O,r)$, we obtain certified robustness properties with respect to l_{p}-perturbations. And by applying to uniform probability measures with support set $B_{\infty}(O,r)$, we obtain certified robustness properties with respect to l_{1},l_{2},l_{∞}-perturbations. We present experimental results on CIFAR-10 dataset with ResNet to validate our theory. It is worth mentioning that our certification procedure only costs constant computation time which is an improvement upon the state-of-the-art methods in terms of the computation time.

****

PRIMA: Planner-Reasoner Inside a Multi-task Reasoning Agent

Daoming Lyu,Bo Liu,Jianshu Chen

In multi-task reasoning (MTR), an agent can solve multiple tasks via (first-order) logic reasoning. This capability is essential for human-like intelligence due to its strong generalizability and simplicity for handling multiple tasks. However, a major challenge in developing effective MTR is the intrinsic conflict between reasoning capability and efficiency. An MTR-capable agent must master a large set of "skills'' to perform diverse tasks, but executing a particular task at the inference stage requires only a small subset of immediately relevant skills. How can we maintain broad reasoning capability yet efficient specific-task performance? To address this problem, we propose a Planner-Reasoner framework capable of state-of-the-art MTR capability and high efficiency. The Reasoner models shareable (first-order) logic deduction rules, from which the Planner selects a subset to compose into efficient reasoning paths. The entire model is trained in an end-to-end manner using deep reinforcement learning, and experimental studies over various domains validate its effectiveness.

****

Zero-Shot Recommender Systems

HAO DING,Yifei Ma,Anoop Deoras,Bernie Wang,Hao Wang

Performance of recommender systems (RecSys) relies heavily on the amount of training data available. This poses a chicken-and-egg problem for early-stage products, whose amount of data, in turn, relies on the performance of their RecSys. In this paper, we explore the possibility of zero-shot learning in RecSys, to enable generalization from an old dataset to an entirely new dataset. We develop an algorithm, dubbed ZEro-Shot Recommenders (ZESRec), that is trained on an old dataset and generalize to a new one where there are neither overlapping users nor overlapping items, a setting that contrasts typical cross-domain RecSys that has either overlapping users or items. Different from previous methods that use categorical item indices (i.e., item ID), ZESRec uses items' generic features, such as natural-language descriptions, product images, and videos, as their continuous indices, and therefore naturally generalizes to any unseen items. In terms of users, ZESRec builds upon recent advances on sequential RecSys to represent users using their interactions with items, thereby generalizing to unseen users as well. We study three pairs of real-world RecSys datasets and demonstrate that ZESRec can successfully enable recommendations in such a zero-shot setting, opening up new opportunities for resolving the chicken-and-egg problem for data-scarce startups or early-stage products.

****

Controlling the Complexity and Lipschitz Constant improves Polynomial Nets

Zhenyu Zhu,Fabian Latorre,Grigorios Chrysos,Volkan Cevher

While the class of Polynomial Nets demonstrates comparable performance to neural networks (NN), it currently has neither theoretical generalization characterization nor robustness guarantees. To this end, we derive new complexity bounds for the set of Coupled CP-Decomposition (CCP) and Nested Coupled CP-decomposition (NCP) models of Polynomial Nets in terms of the ℓ_∞-operator-norm and the ℓ_2-operator norm. In addition, we derive bounds on the Lipschitz constant for both models to establish a theoretical certificate for their robustness.

The theoretical results enable us to propose a principled regularization scheme that we also evaluate experimentally and show that it improves the accuracy as well as the robustness of the models to adversarial perturbations. We showcase how this regularization can be combined with adversarial training, resulting in further improvements.

****

Open-Set Recognition: A Good Closed-Set Classifier is All You Need

Sagar Vaze,Kai Han,Andrea Vedaldi,Andrew Zisserman

The ability to identify whether or not a test sample belongs to one of the semantic classes in a classifier's training set is critical to practical deployment of the model. This task is termed open-set recognition (OSR) and has received significant attention in recent years. In this paper, we first demonstrate that the ability of a classifier to make the 'none-of-above' decision is highly correlated with its accuracy on the closed-set classes. We find that this relationship holds across loss objectives and architectures, and further demonstrate the trend both on the standard OSR benchmarks as well as on a large-scale ImageNet evaluation. Second, we use this correlation to boost the performance of the maximum softmax probability OSR 'baseline' by improving its closed-set accuracy, and with this strong baseline achieve state-of-the-art on a number of OSR benchmarks. Similarly, we boost the performance of the existing state-of-the-art method by improving its closed-set accuracy, but the resulting discrepancy with the strong baseline is marginal. Our third contribution is to present the 'Semantic Shift Benchmark' (SSB), which better respects the task of detecting semantic novelty, as opposed to low-level distributional shifts as tackled by neighbouring machine learning fields. On this new evaluation, we again demonstrate that there is negligible difference between the strong baseline and the existing state-of-the-art. Code available at: https://github.com/sgvaze/osr_closed_set_all_you_need.

****

That Escalated Quickly: Compounding Complexity by Editing Levels at the Frontier of Agent Capabilities

Jack Parker-Holder,Minqi Jiang,Michael D Dennis,Mikayel Samvelyan,Jakob Nicolaus Foerster,Edward Grefenstette,Tim Rocktäschel

Deep Reinforcement Learning (RL) has recently produced impressive results in a series of settings such as games and robotics. However, a key challenge that limits the utility of RL agents for real-world problems is the agent's ability to generalize to unseen variations (or levels). To train more robust agents, the field of Unsupervised Environment Design (UED) seeks to produce a curriculum by updating both the agent and the distribution over training environments. Recent advances in UED have come from promoting levels with high regret, which provides theoretical guarantees in equilibrium and empirically has been shown to produce agents capable of zero-shot transfer to unseen human-designed environments. However, current methods require either learning an environment-generating adversary, which remains a challenging optimization problem, or curating a curriculum from randomly sampled levels, which is ineffective if the search space is too large. In this paper we instead propose to evolve a curriculum, by making edits to previously selected levels. Our approach, which we call Adversarially Compounding Complexity by Editing Levels (ACCEL), produces levels at the frontier of an agent's capabilities, resulting in curricula that start simple but become increasingly complex. ACCEL maintains the theoretical benefits of prior works, while outperforming them empirically when transferring to complex out-of-distribution environments.

****

Finding an Unsupervised Image Segmenter in each of your Deep Generative Models

Luke Melas-Kyriazi,Christian Rupprecht,Iro Laina,Andrea Vedaldi

Recent research has shown that numerous human-interpretable directions exist in the latent space of GANs. In this paper, we develop an automatic procedure for finding directions that lead to foreground-background image separation, and we use these directions to train an image segmentation model without human supervision. Our method is generator-agnostic, producing strong segmentation results with a wide range of different GAN architectures. Furthermore, by leveraging GANs pre trained on large datasets such as ImageNet, we are able to segment images from a range of domains without further training or finetuning. Evaluating our method on image segmentation benchmarks, we compare favorably to prior work while using neither human supervision nor access to the training data. Broadly, our results demonstrate that automatically extracting foreground-background structure from pretrained deep generative models can serve as a remarkably effective substitute for human supervision.

****

Interest-based Item Representation Framework for Recommendation with Multi-Interests Capsule Network

Yanpeng Xie,Tong Zhang,Heng Zhang,Zhendong Qu

Item representation plays an important role for recommendation, such as e-commerce, news, video, etc. It has been used by retrieval and ranking model to capture user-item relationship based on user behaviors. For recommendation systems, user interaction behaviors imply single or multi interests of the user, not only items themselves in the sequences. Existing representation learning methods mainly focus on optimizing item-based mechanism between user interaction sequences and candidate item(especially attention mechanism, sequential modeling). However, item representations learned by these methods lack modeling mechanism to reflect user interests. That is, the methods may be less effective and indirect to capture user interests. We propose a framework to learn interest-based item representations directly by introducing user Multi Interests Capsule Network(MICN). To make the framework model-agnostic, user Multi Interests Capsule Network is designed as an auxiliary task to jointly learn item-based item representations and interest-based item representations. Hence, the generic framework can be easily used to improve existing recommendation models without model redesign. The proposed approach is evaluated on multiple types of benchmarks. Furthermore, we investigate several situations on various deep neural networks, different length of behavior sequences and joint learning ratio of interest-based item representations. Experiment shows a great enhancement on performance of various recommendation models and has also validated our approach. We expect the framework could be widely used for recommendation systems.

****

Token Pooling in Vision Transformers

Dmitrii Marin,Jen-Hao Rick Chang,Anurag Ranjan,Anish Prabhu,Mohammad Rastegari,Oncel Tuzel

Despite the recent success in many applications, the high computational requirements of vision transformers limit their use in resource-constrained settings. While many existing methods improve the quadratic complexity of attention, in most vision transformers, self-attention is not the major computation bottleneck, e.g., more than 80% of the computation is spent on fully-connected layers. To improve the computational complexity of all layers, we propose a novel token downsampling method, called Token Pooling, efficiently exploiting redundancies in the images and intermediate token representations. We show that, under mild assumptions, softmax-attention acts as a high-dimensional low-pass (smoothing) filter. Thus, its output contains redundancy that can be pruned to achieve a better trade-off between the computational cost and accuracy. Our new technique accurately approximates a set of tokens by minimizing the reconstruction error caused by downsampling. We solve this optimization problem via cost-efficient clustering. We rigorously analyze and compare to prior downsampling methods. Our experiments show that Token Pooling significantly improves the cost-accuracy trade-off over the state-of-the-art downsampling. Token Pooling is a simple and effective operator that can benefit many architectures. Applied to DeiT, it achieves the same ImageNet top-1 accuracy using 42% fewer computations.

****

Reducing the Teacher-Student Gap via Adaptive Temperatures

Jia Guo

Knowledge distillation aims to obtain a small and effective deep model (student) by learning the output from a larger model (teacher). Previous studies found a severe degradation problem, that student performance would degrade unexpectedly when distilled from oversized teachers. It is well known that larger models tend to have sharper outputs. Based on this observation, we found that the sharpness gap between the teacher and student output may cause this degradation problem. To solve this problem, we first propose a metric to quantify the sharpness of the model output. Based on the second-order Taylor expansion of this metric, we propose Adaptive Temperature Knowledge Distillation (ATKD), which automatically changes the temperature of the teacher and the student, to reduce the sharpness gap. We conducted extensive experiments on CIFAR100 and ImageNet and achieved significant improvements. Specifically, ATKD trained the best ResNet18 model on ImageNet as we knew (73.0% accuracy).

****

Mako: Semi-supervised continual learning with minimal labeled data via data programming

Pengyuan Lu,Seungwon Lee,Amanda Watson,David Kent,Insup Lee,ERIC EATON,James Weimer

Lifelong machine learning (LML) is a well-known paradigm mimicking the human learning process by utilizing experiences from previous tasks. Nevertheless, an issue that has been rarely addressed is the lack of labels at the individual task level. The state-of-the-art of LML largely addresses supervised learning, with a few semi-supervised continual learning exceptions which require training additional models, which in turn impose constraints on the LML methods themselves. Therefore, we propose Mako, a wrapper tool that mounts on top of supervised LML frameworks, leveraging data programming. Mako imposes no additional knowledge base overhead and enables continual semi-supervised learning with a limited amount of labeled data. This tool achieves similar performance, in terms of per-task accuracy and resistance to catastrophic forgetting, as compared to fully labeled data. We ran extensive experiments on LML task sequences created from standard image classification data sets including MNIST, CIFAR-10 and CIFAR-100, and the results show that after utilizing Mako to leverage unlabeled data, LML tools are able to achieve 97% performance of supervised learning on fully labeled data in terms of accuracy and catastrophic forgetting prevention. Moreover, when compared to baseline semi-supervised LML tools such as CNNL, ORDisCo and DistillMatch, Mako significantly outperforms them, increasing accuracy by 0.25 on certain benchmarks.

****

Solving Inverse Problems in Medical Imaging with Score-Based Generative Models

Yang Song,Liyue Shen,Lei Xing,Stefano Ermon

Reconstructing medical images from partial measurements is an important inverse problem in Computed Tomography (CT) and Magnetic Resonance Imaging (MRI). Existing solutions based on machine learning typically train a model to directly map measurements to medical images, leveraging a training dataset of paired images and measurements. These measurements are typically synthesized from images using a fixed physical model of the measurement process, which hinders the generalization capability of models to unknown measurement processes. To address this issue, we propose a fully unsupervised technique for inverse problem solving, leveraging the recently introduced score-based generative models. Specifically, we first train a score-based generative model on medical images to capture their prior distribution. Given measurements and a physical model of the measurement process at test time, we introduce a sampling method to reconstruct an image consistent with both the prior and the observed measurements. Our method does not assume a fixed measurement process during training, and can thus be flexibly adapted to different measurement processes at test time. Empirically, we observe comparable or better performance to supervised learning techniques in several medical imaging tasks in CT and MRI, while demonstrating significantly better generalization to unknown measurement processes.

****

Scalable One-Pass Optimisation of High-Dimensional Weight-Update Hyperparameters by Implicit Differentiation

Ross M Clarke,Elre Talea Oldewage,José Miguel Hernández-Lobato

Machine learning training methods depend plentifully and intricately on hyperpar ameters, motivating automated strategies for their optimisation. Many existing algorithms restart training for each new hyperparameter choice, at considerable computational cost. Some hypergradient-based one-pass methods exist, but these either cannot be applied to arbitrary optimiser hyperparameters (such as learning rates and momenta) or take several times longer to train than their base models. We extend these existing methods to develop an approximate hypergradient-based hyperparameter optimiser which is applicable to any continuous hyperparameter appearing in a differentiable model weight update, yet requires only one training episode, with no restarts. We also provide a motivating argument for convergence to the true hypergradient, and perform tractable gradient-based optimisation of independent learning rates for each model parameter. Our method performs competitively from varied random hyperparameter initialisations on several UCI datasets and Fashion-MNIST (using a one-layer MLP), Penn Treebank (using an LSTM) and CIFAR-10 (using a ResNet-18), in time only 2-3x greater than vanilla training.

****

Exploring Covariate and Concept Shift for Detection and Confidence Calibration of Out-of-Distribution Data

Junjiao Tian,Yen-Chang Hsu,Yilin Shen,Hongxia Jin,Zsolt Kira

Moving beyond testing on in-distribution data, works on Out-of-Distribution (OOD) detection have recently increased in popularity. A recent attempt to categorize OOD data introduces the concept of near and far OOD detection. Specifically, prior works define characteristics of OOD data in terms of detection difficulty. We propose to characterize the spectrum of OOD data using two types of distribution shifts: covariate shift and concept shift, where covariate shift corresponds to change in style, e.g., noise, and concept shift indicates change in semantics. This characterization reveals that sensitivity to each type of shift is important to the detection and model calibration of OOD data. Consequently, we investigate score functions that capture sensitivity to each type of dataset shift and methods that improve them. To this end, we theoretically derive two score functions for OOD detection, the covariate shift score and concept shift score, based on the decomposition of KL-divergence for both scores, and propose a geometrically-inspired method (Geometric ODIN) to improve OOD detection under both shifts with only in-distribution data. Additionally, the proposed method naturally leads to an expressive post-hoc calibration function which yields state-of-the-art calibration performance on both in-distribution and out-of-distribution data. We are the first to propose a method that works well across both OOD detection and calibration, and under different types of shifts. Specifically, we improve the previous state-of-the-art OOD detection by relatively 7% AUROC on CIFAR100 vs. SVHN and achieve the best calibration performance of 0.084 Expected Calibration Error on the corrupted CIFAR100C dataset.

****

Sample Efficient Deep Reinforcement Learning via Uncertainty Estimation

Vincent Mai,Kaustubh Mani,Liam Paull

In model-free deep reinforcement learning (RL) algorithms, using noisy value estimates to supervise policy evaluation and optimization is detrimental to the sample efficiency. As this noise is heteroscedastic, its effects can be mitigated using uncertainty-based weights in the optimization process. Previous methods rely on sampled ensembles, which do not capture all aspects of uncertainty. We provide a systematic analysis of the sources of uncertainty in the noisy supervision that occurs in RL, and introduce inverse-variance RL, a Bayesian framework which combines probabilistic ensembles and Batch Inverse Variance weighting. We propose a method whereby two complementary uncertainty estimation methods account for both the Q-value and the environment stochasticity to better mitigate the negative impacts of noisy supervision. Our results show significant improvement in terms of sample efficiency on discrete and continuous control tasks.

****

Greedy-based Value Representation for Efficient Coordination in Multi-agent Reinforcement Learning

Lipeng Wan,Zeyang Liu,Xingyu Chen,Han Wang,Xuguang Lan

Due to the representation limitation of the joint Q value function, multi-agent

reinforcement learning (MARL) methods with linear or monotonic value decomposition can not ensure the optimal consistency (i.e. the correspondence between the individual greedy actions and the maximal true Q value), leading to instability and poor coordination. Existing methods focus on addressing the representation limitation through learning the complete expressiveness, which is impractical and may deteriorate the performance in complex tasks. In this paper, we introduce the True-Global-Max (TGM) condition for linear and monotonic value decomposition to achieve the optimal consistency directly, where the TGM condition can be ensured under the unique stability of the optimal greedy action. Therefore, we propose the greedy-based value representation (GVR), which stabilises the optimal greedy action via inferior target shaping and destabilises non-optimal greedy actions via superior experience replay. We conduct experiments on various benchmarks, where GVR significantly outperforms state-of-the-art baselines. Experiment results demonstrate that our method can meet the optimal consistency under sufficient exploration and is more efficient than methods with complete expressiveness capability.
****

An Empirical Study of Pre-trained Models on Out-of-distribution Generalization
Yaodong Yu,Heinrich Jiang,Dara Bahri,Hossein Mobahi,Seungyeon Kim,Ankit Singh Rawat,Andreas Veit,Yi Ma
Generalizing to out-of-distribution (OOD) data -- that is, data from domains unseen during training -- is a key challenge in modern machine learning, which has only recently received much attention. Some existing approaches propose leveraging larger models and pre-training on larger datasets. In this paper, we provide new insights in applying these approaches. Concretely, we show that larger models and larger datasets need to be simultaneously leveraged to improve OOD performance. Moreover, we show that using smaller learning rates during fine-tuning is critical to achieving good results, contrary to popular intuition that larger learning rates generalize better when training from scratch. We show that strategies that improve in-distribution accuracy may, counter-intuitively, lead to poor OOD performance despite strong in-distribution performance. Our insights culminate to a method that achieves state-of-the-art results on a number of OOD generalization benchmark tasks, often by a significant margin.
****

Contrastive Pre-training for Zero-Shot Information Retrieval
Gautier Izacard,Mathilde Caron,Lucas Hosseini,Sebastian Riedel,Piotr Bojanowski,Armand Joulin,Edouard Grave
Information retrieval is an important component in natural language processing, for knowledge intensive tasks such as question answering and fact checking. Recently, information retrieval has seen the emergence of dense retrievers, based on neural networks, as an alternative to classical sparse methods based on term-frequency. Neural retrievers work well on the problems for which they were specifically trained, but they do not generalize as well as term-frequency methods to new domains or applications. By contrast, in many other NLP tasks, conventional self-supervised pre-training based on masking leads to strong generalization with small number of training examples. We believe this is not yet the case for information retrieval, because these pre-training methods are not well adapted to this task. In this work, we consider contrastive learning as a more natural pre-training technique for retrieval and show that it leads to models that are competitive with BM25 on many domains or applications, even without training on supervised data. Our dense pre-trained models also compare favorably against BERT pre-trained models in the few-shot setting, and achieves state-of-the-art performance on the BEIR benchmark when fine-tuned on MS-MARCO.
****

Learning Pruning-Friendly Networks via Frank-Wolfe: One-Shot, Any-Sparsity, And No Retraining
Lu Miao,Xiaolong Luo,Tianlong Chen,Wuyang Chen,Dong Liu,Zhangyang Wang
We present a novel framework to train a large deep neural network (DNN) for only $\textit{once}$, which can then be pruned to $\textit{any sparsity ratio}$ to preserve competitive accuracy $\textit{without any re-training}$. Conventional me

thods often require (iterative) pruning followed by re-training, which not only incurs large overhead beyond the original DNN training but also can be sensitive to retraining hyperparameters. Our core idea is to re-cast the DNN training as an explicit $\textit{pruning-aware}$ process: that is formulated with an auxiliary K-sparse polytope constraint, to encourage network weights to lie in a convex hull spanned by K-sparse vectors, potentially resulting in more sparse weight matrices. We then leverage a stochastic Frank-Wolfe (SFW) algorithm to solve this new constrained optimization, which naturally leads to sparse weight updates each time. We further note an overlooked fact that existing DNN initializations were derived to enhance SGD training (e.g., avoid gradient explosion or collapse), but was unaligned with the challenges of training with SFW. We hence also present the first learning-based initialization scheme specifically for boosting SFW-based DNN training. Experiments on CIFAR-10 and Tiny-ImageNet datasets demonstrate that our new framework named $\textbf{SFW-pruning}$ consistently achieves the state-of-the-art performance on various benchmark DNNs over a wide range of pruning ratios. Moreover, SFW-pruning only needs to train once on the same model and dataset, for obtaining arbitrary ratios, while requiring neither iterative pruning nor retraining. All codes will be released to the public.
****
BDDM: Bilateral Denoising Diffusion Models for Fast and High-Quality Speech Synthesis
Max W. Y. Lam,Jun Wang,Dan Su,Dong Yu
Diffusion probabilistic models (DPMs) and their extensions have emerged as competitive generative models yet confront challenges of efficient sampling. We propose a new bilateral denoising diffusion model (BDDM) that parameterizes both the forward and reverse processes with a schedule network and a score network, which can train with a novel bilateral modeling objective. We show that the new surrogate objective can achieve a lower bound of the log marginal likelihood tighter than a conventional surrogate. We also find that BDDM allows inheriting pre-trained score network parameters from any DPMs and consequently enables speedy and stable learning of the schedule network and optimization of a noise schedule for sampling.
Our experiments demonstrate that BDDMs can generate high-fidelity audio samples with as few as three sampling steps. Moreover, compared to other state-of-the-art diffusion-based neural vocoders, BDDMs produce comparable or higher quality samples indistinguishable from human speech, notably with only seven sampling steps (143x faster than WaveGrad and 28.6x faster than DiffWave). We release our code at https://github.com/tencent-ailab/bddm.
****
Learning Structure from the Ground up---Hierarchical Representation Learning by Chunking
Shuchen Wu,Noemi Elteto,Ishita Dasgupta,Eric Schulz
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 representation 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 visual, temporal, visual-temporal domains and language data. Furthermore, the interpretability of the learned chunks enables flexible transfer between environments that share partial representational structure. Taken together, our results show how cognitive science in general and theories of chunking in particular could inform novel and more interpretable approaches to representation learning.
****
Understanding Overfitting in Reweighting Algorithms for Worst-group Performance

Runtian Zhai,Chen Dan,J Zico Kolter,Pradeep Kumar Ravikumar

Prior work has proposed various reweighting algorithms to improve the worst-group performance of machine learning models for fairness. However, Sagawa et al. (2020) empirically found that these algorithms overfit easily in practice under the overparameterized setting, where the number of model parameters is much greater than the number of samples. In this work, we provide a theoretical backing to the empirical results above, and prove the pessimistic result that reweighting algorithms always overfit. Specifically we prove that with reweighting, an overparameterized model always converges to the same ERM interpolator that fits all training samples, and consequently its worst-group test performance will drop to the same level as ERM in the long run. That is, we cannot hope for reweighting algorithms to converge to a different interpolator than ERM with potentially better worst-group performance. Then, we analyze whether adding regularization helps fix the issue, and we prove that for regularization to work, it must be large enough to prevent the model from achieving small training error. Our results suggest that large regularization (or early stopping) and data augmentation are necessary for reweighting algorithms to achieve high worst-group test performance.

****

ED2: An Environment Dynamics Decomposition Framework for World Model Construction

Cong Wang,Tianpei Yang,Jianye HAO,YAN ZHENG,Hongyao Tang,Fazl Barez,Jinyi Liu,Jiajie Peng,haiyin piao,Zhixiao Sun

Model-based reinforcement learning methods achieve significant sample efficiency in many tasks, but their performance is often limited by the existence of the model error. To reduce the model error, previous works use a single well-designed network to fit the entire environment dynamics, which treats the environment dynamics as a black box. However, these methods lack to consider the environmental decomposed property that the dynamics may contain multiple sub-dynamics, which can be modeled separately, allowing us to construct the world model more accurately. In this paper, we propose the Environment Dynamics Decomposition (ED2), a novel world model construction framework that models the environment in a decomposing manner. ED2 contains two key components: sub-dynamics discovery (SD2) and dynamics decomposition prediction (D2P). SD2 discovers the sub-dynamics in an environment and then D2P constructs the decomposed world model following the sub-dynamics. ED2 can be easily combined with existing MBRL algorithms and empirical results show that ED2 significantly reduces the model error and boosts the performance of the state-of-the-art MBRL algorithms on various continuous control tasks.

****

How to measure deep uncertainty estimation performance and which models are naturally better at providing it

Ido Galil,Mohammed Dabbah,Ran El-Yaniv

When deployed for risk-sensitive tasks, deep neural networks (DNNs) must be equipped with an uncertainty estimation mechanism. This paper studies the relationship between deep architectures and their training regimes with their corresponding uncertainty estimation performance. We consider both in-distribution uncertainties ("aleatoric" or "epistemic") and class-out-of-distribution ones. Moreover, we consider some of the most popular estimation performance metrics previously proposed including AUROC, ECE, AURC, and coverage for selective accuracy constraint. We present a novel and comprehensive study carried out by evaluating the uncertainty performance of 484 deep ImageNet classification models.

We identify numerous and previously unknown factors that affect uncertainty estimation and examine the relationships between the different metrics. We find that distillation-based training regimes consistently yield better uncertainty estimations than other training schemes such as vanilla training, pretraining on a larger dataset and adversarial training. We also provide strong empirical evidence showing that ViT is by far the most superior architecture in terms of uncertainty estimation performance, judging by any aspect, in both in-distribution and class-out-of-distribution scenarios. We learn various interesting facts along the way. Contrary to previous work, ECE does not necessarily worsen with an increase

in the number of network parameters. Likewise, we discovered an unprecedented 99% top-1 selective accuracy at 47% coverage (and 95% top-1 accuracy at 80%) for a ViT model, whereas a competing EfficientNet-V2-XL cannot obtain these accuracy constraints at any level of coverage.
****

Sample Efficient Stochastic Policy Extragradient Algorithm for Zero-Sum Markov Game

Ziyi Chen,Shaocong Ma,Yi Zhou

Two-player zero-sum Markov game is a fundamental problem in reinforcement learning and game theory. Although many algorithms have been proposed for solving zero-sum Markov games in the existing literature, many of them either require a full knowledge of the environment or are not sample-efficient. In this paper, we develop a fully decentralized and sample-efficient stochastic policy extragradient algorithm for solving tabular zero-sum Markov games. In particular, our algorithm utilizes multiple stochastic estimators to accurately estimate the value functions involved in the stochastic updates, and leverages entropy regularization to accelerate the convergence. Specifically, with a proper entropy-regularization parameter, we prove that the stochastic policy extragradient algorithm has a sample complexity of the order $\widetilde{\mathcal{O}}(\frac{A_{\max}}{\mu_{\text{min}}\epsilon^{5.5}(1-\gamma)^{13.5}})$ for finding a solution that achieves ϵ-Nash equilibrium duality gap, where $A_{\max}$ is the maximum number of actions between the players, $\mu_{\min}$ is the lower bound of state stationary distribution, and γ is the discount factor. Such a sample complexity result substantially improves the state-of-the-art complexity result.
****

Learning transferable motor skills with hierarchical latent mixture policies

Dushyant Rao,Fereshteh Sadeghi,Leonard Hasenclever,Markus Wulfmeier,Martina Zambelli,Giulia Vezzani,Dhruva Tirumala,Yusuf Aytar,Josh Merel,Nicolas Heess,raia hadsell

For robots operating in the real world, it is desirable to learn reusable abstract behaviours that can effectively be transferred across numerous tasks and scenarios.

We propose an approach to learn skills from data using a hierarchical mixture latent variable model.

Our method exploits a multi-level hierarchy of both discrete and continuous latent variables, to model a discrete set of abstract high-level behaviours while allowing for variance in how they are executed.

We demonstrate in manipulation domains that the method can effectively cluster offline data into distinct, executable behaviours, while retaining the flexibility of a continuous latent variable model.

The resulting skills can be transferred to new tasks, unseen objects, and from state to vision-based policies, yielding significantly better sample efficiency and asymptotic performance compared to existing skill- and imitation-based methods.

We also perform further analysis showing how and when the skills are most beneficial: they encourage directed exploration to cover large regions of the state space relevant to the task, making them most effective in challenging sparse-reward settings.
****

Zeroth-Order Actor-Critic

Yuheng Lei,Jianyu Chen,Shengbo Eben Li,Sifa Zheng

Evolution based zeroth-order optimization methods and policy gradient based first-order methods are two promising alternatives to solve reinforcement learning (RL) problems with complementary advantages. The former work with arbitrary policies, drive state-dependent and temporally-extended exploration, possess robustness-seeking property, but suffer from high sample complexity, while the latter are more sample efficient but restricted to differentiable policies and the learned policies are less robust. We propose Zeroth-Order Actor-Critic algorithm (ZOAC) that unifies these two methods into an on-policy actor-critic architecture to preserve the advantages from both. ZOAC conducts rollouts collection with timest ep-wise perturbation in parameter space, first-order policy evaluation (PEV) and zeroth-order policy improvement (PIM) alternately in each iteration. The modified rollouts collection strategy and the introduced critic network help to reduce the variance of zeroth-order gradient estimators and improve the sample efficiency and stability of the algorithm. We evaluate our proposed method using two different types of policies, linear policies and neural networks, on a range of challenging continuous control benchmarks, where ZOAC outperforms zeroth-order and first-order baseline algorithms.

****

The Uncanny Similarity of Recurrence and Depth

Avi Schwarzschild,Arjun Gupta,Amin Ghiasi,Micah Goldblum,Tom Goldstein

It is widely believed that deep neural networks contain layer specialization, wherein networks extract hierarchical features representing edges and patterns in shallow layers and complete objects in deeper layers. Unlike common feed-forward models that have distinct filters at each layer, recurrent networks reuse the same parameters at various depths. In this work, we observe that recurrent models exhibit the same hierarchical behaviors and the same performance benefits as depth despite reusing the same filters at every recurrence. By training models of various feed-forward and recurrent architectures on several datasets for image classification as well as maze solving, we show that recurrent networks have the ability to closely emulate the behavior of non-recurrent deep models, often doing so with far fewer parameters.

****

Surgical Prediction with Interpretable Latent Representation

Bing Xue,York Jiao,Thomas Kannampallil,Joanna Abraham,Christopher Ryan King,Bradley A Fritz,Michael Avidan,Chenyang Lu

Given the risks and cost of surgeries, there has been significant interest in exploiting predictive models to improve perioperative care. However, due to the high dimensionality and noisiness of perioperative data, it is challenging to develop accurate, robust and interpretable encoding for surgical applications. We propose surgical VAE (sVAE), a representation learning framework for perioperative data based on variational autoencoder (VAE). sVAE provides a holistic approach combining two salient features tailored for surgical applications. To overcome performance limitations of traditional VAE, it is prediction-guided with explicit expression of predicted outcome in the latent representation. Furthermore, it disentangles the latent space so that it can be interpreted in a clinically meaningful fashion. We apply sVAE to two real-world perioperative datasets and the open MIMIC-III dataset to evaluate its efficacy and performance in predicting diverse outcomes including surgery duration, postoperative complication, ICU duration, and mortality. Our results show that the latent representation provided by sVAE leads to superior performance in classification, regression and multi-task predictions. We further demonstrate the interpretability of the disentangled representation and its capability to capture intrinsic characteristics of surgical patients.

****

Implicit Bias of Adversarial Training for Deep Neural Networks

Bochen Lv,Zhanxing Zhu

We provide theoretical understandings of the implicit bias imposed by adversarial training for homogeneous deep neural networks without any explicit regularization. In particular, for deep linear networks adversarially trained by gradient descent on a linearly separable dataset, we prove that the direction of the product of weight matrices converges to the direction of the max-margin solution of the original dataset. Furthermore, we generalize this result to the case of adversarial training for non-linear homogeneous deep neural networks without the linear separability of the dataset. We show that, when the neural network is adversarially trained with ℓ_2 or ℓ_{∞} FGSM, FGM and PGD perturbations, the direction of the limit point of normalized parameters of the network along the trajectory of the gradient flow converges to a KKT point of a constrained optimization problem that aims to maximize the margin for adversarial examples. Our results theoretically justify the longstanding conjecture that adversarial tr aining modifies the decision boundary by utilizing adversarial examples to improve robustness, and potentially provides insights for designing new robust training strategies.

****

Compositional Training for End-to-End Deep AUC Maximization

Zhuoning Yuan,Zhishuai Guo,Nitesh Chawla,Tianbao Yang

Recently, deep AUC maximization (DAM) has achieved great success in different domains (e.g., medical image classification). However, the end-to-end training for deep AUC maximization still remains a challenging problem. Previous studies employ an ad-hoc two-stage approach that first trains the network by optimizing a traditional loss (e.g., cross-entropy loss) and then finetunes the network by optimizing an AUC loss. This is because that training a deep neural network from scratch by maximizing an AUC loss usually does not yield a satisfactory performance. This phenomenon can be attributed to the degraded feature representations learned by maximizing the AUC loss from scratch. To address this issue, we propose a novel compositional training framework for end-to-end DAM, namely compositional DAM. The key idea of compositional training is to minimize a compositional objective function, where the outer function corresponds to an AUC loss and the inner function represents a gradient descent step for minimizing a traditional loss, e.g., the cross-entropy (CE) loss. To optimize the non-standard compositional objective, we propose an efficient and provable stochastic optimization algorithm. The proposed algorithm enhances the capabilities of both robust feature learning and robust classifier learning by alternatively taking a gradient descent step for the CE loss and for the AUC loss in a systematic way. We conduct extensive empirical studies on imbalanced benchmark and medical image datasets, which unanimously verify the effectiveness of the proposed method. Our results show that the compositional training approach dramatically improves both the feature representations and the testing AUC score compared with traditional deep learning approaches, and yields better performance than the two-stage approaches for DAM as well. The proposed method is implemented in our open-sourced library LibAUC (https://www.libauc.org) and code is available at https://github.com/Optimization-AI/LibAUC.

****

Unsupervised Object Learning via Common Fate

Matthias Tangemann,Steffen Schneider,Julius Von Kügelgen,Francesco Locatello,Peter Vincent Gehler,Thomas Brox,Matthias Kuemmerer,Matthias Bethge,Bernhard Schölkopf

Learning generative object models from unlabelled videos is a long standing problem and is required for causal scene modeling. We decompose this problem into three easier subtasks, and provide candidate solutions for each of them. Inspired by the Common Fate Principle of Gestalt Psychology, we first extract (noisy) masks of moving objects via unsupervised motion segmentation. Second, generative models are trained on the masks of the background and the moving objects, respectively. Third, background and foreground models are combined in a conditional ``dead leaves scene model to sample novel scene configurations where occlusions and depth layering arise naturally. To evaluate the individual stages, we introduce the Fishbowl dataset positioned between complex real-world scenes and common object-centric benchmarks of simplistic objects. We show that our approach allows learning generative models that generalize beyond the occlusions present in the input videos, and represent scenes in a modular fashion that allows sampling plausible scenes outside the training distribution by permitting, for instance, object numbers or densities not observed in the training set.

****

Explanations of Black-Box Models based on Directional Feature Interactions

Aria Masoomi,Davin Hill,Zhonghui Xu,Craig P Hersh,Edwin K. Silverman,Peter J. Castaldi,Stratis Ioannidis,Jennifer Dy

As machine learning algorithms are deployed ubiquitously to a variety of domains, it is imperative to make these often black-box models transparent. Several recent works explain black-box models by capturing the most influential features for prediction per instance; such explanation methods are univariate, as they cha racterize importance per feature. We extend univariate explanation to a higher-order; this enhances explainability, as bivariate methods can capture feature interactions in black-box models, represented as a directed graph. Analyzing this graph enables us to discover groups of features that are equally important (i.e., interchangeable), while the notion of directionality allows us to identify the most influential features. We apply our bivariate method on Shapley value explanations, and experimentally demonstrate the ability of directional explanations to discover feature interactions. We show the superiority of our method against state-of-the-art on CIFAR10, IMDB, Census, Divorce, Drug, and gene data.

****

Mix-MaxEnt: Creating High Entropy Barriers To Improve Accuracy and Uncertainty Estimates of Deterministic Neural Networks

Francesco Pinto,Harry Yang,Ser-Nam Lim,Philip Torr,Puneet K. Dokania

We propose an extremely simple approach to regularize a single deterministic neural network to obtain improved accuracy and reliable uncertainty estimates. Our approach, on top of the cross-entropy loss, simply puts an entropy maximization regularizer corresponding to the predictive distribution in the regions of the embedding space between the class clusters. This is achieved by synthetically generating between-cluster samples via the convex combination of two images from different classes and maximizing the entropy on these samples. Such a data-dependent regularization guides the maximum likelihood estimation to prefer a solution that (1) maps out-of-distribution samples to high entropy regions (creating an entropy barrier); and (2) is more robust superficial input perturbations.

Via extensive experiments on real-world datasets (CIFAR-10 and CIFAR-100) using ResNet and Wide-ResNet architectures, we demonstrate that Mix-MaxEnt consistently provides much improved classification accuracy, better calibrated probabilities for in-distribution data, and reliable uncertainty estimates when exposed to situations involving domain-shift and out-of-distribution samples.

****

Surprise Minimizing Multi-Agent Learning with Energy-based Models

Karush Suri

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.

****

Mastering Visual Continuous Control: Improved Data-Augmented Reinforcement Learning

Denis Yarats,Rob Fergus,Alessandro Lazaric,Lerrel Pinto

We present DrQ-v2, a model-free reinforcement learning (RL) algorithm for visual continuous control. DrQ-v2 builds on DrQ, an off-policy actor-critic approach that uses data augmentation to learn directly from pixels. We introduce several improvements that yield state-of-the-art results on the DeepMind Control Suite. Notably, DrQ-v2 is able to solve complex humanoid locomotion tasks directly from pixel observations, previously unattained by model-free RL. DrQ-v2 is conceptually simple, easy to implement, and provides significantly better computational footprint compared to prior work, with the majority of tasks taking just 8 hours to train on a single GPU. Finally, we publicly release DrQ-v2 's implementation to provide RL practitioners with a strong and computationally efficient baselin e.

****

Regularizing Deep Neural Networks with Stochastic Estimators of Hessian Trace

Yucong Liu,Tong Lin

In this paper we develop a novel regularization method for deep neural networks by penalizing the trace of Hessian. This regularizer is motivated by a recent guarantee bound of the generalization error. Hutchinson method is a classical unbiased estimator for the trace of a matrix, but it is very time-consuming on deep learning models. Hence a dropout scheme is proposed to efficiently implements the Hutchinson method. Then we discuss a connection to linear stability of a nonlinear dynamical system. Experiments demonstrate that our method outperforms existing regularizers such as Jacobian, confidence penalty, and label smoothing. Our regularization method is also orthogonal to data augmentation methods, achieving the best performance when our method is combined with data augmentation.

****

πBO: Augmenting Acquisition Functions with User Beliefs for Bayesian Optimization

Carl Hvarfner,Danny Stoll,Artur Souza,Marius Lindauer,Frank Hutter,Luigi Nardi

Bayesian optimization (BO) has become an established framework and popular tool for hyperparameter optimization (HPO) of machine learning (ML) algorithms. While known for its sample-efficiency, vanilla BO can not utilize readily available prior beliefs the practitioner has on the potential location of the optimum. Thus, BO disregards a valuable source of information, reducing its appeal to ML practitioners. To address this issue, we propose πBO, an acquisition function generalization which incorporates prior beliefs about the location of the optimum in the form of a probability distribution, provided by the user. In contrast to previous approaches, πBO is conceptually simple and can easily be integrated with existing libraries and many acquisition functions. We provide regret bounds when πBO is applied to the common Expected Improvement acquisition function and prove convergence at regular rates independently of the prior. Further, our experiments show that πBO outperforms competing approaches across a wide suite of benchmarks and prior characteristics. We also demonstrate that πBO improves on the state-of-the-art performance for a popular deep learning task, with a $12.5\times$ time-to-accuracy speedup over prominent BO approaches.

****

A Generalized Weighted Optimization Method for Computational Learning and Inversion

Kui Ren,Yunan Yang,Björn Engquist

The generalization capacity of various machine learning models exhibits different phenomena in the under- and over-parameterized regimes. In this paper, we focus on regression models such as feature regression and kernel regression and analyze a generalized weighted least-squares optimization method for computational learning and inversion with noisy data. The highlight of the proposed framework is that we allow weighting in both the parameter space and the data space. The weighting scheme encodes both a priori knowledge on the object to be learned and a strategy to weight the contribution of different data points in the loss function. Here, we characterize the impact of the weighting scheme on the generalization error of the learning method, where we derive explicit generalization errors for the random Fourier feature model in both the under- and over-parameterized regimes. For more general feature maps, error bounds are provided based on the singular values of the feature matrix. We demonstrate that appropriate weighting from prior knowledge can improve the generalization capability of the learned model.

****

Learning Strides in Convolutional Neural Networks

Rachid Riad,Olivier Teboul,David Grangier,Neil Zeghidour

Convolutional neural networks typically contain several downsampling operators, such as strided convolutions or pooling layers, that progressively reduce the resolution of intermediate representations. This provides some shift-invariance while reducing the computational complexity of the whole architecture. A critical hyperparameter of such layers is their stride: the integer factor of downsampling. As strides are not differentiable, finding the best configuration either requires cross-validation or discrete optimization (e.g. architecture search), which rapidly become prohibitive as the search space grows exponentially with the number of downsampling layers. Hence, exploring this search space by gradient descent would allow finding better configurations at a lower computational cost. This work introduces DiffStride, the first downsampling layer with learnable strides. Our layer learns the size of a cropping mask in the Fourier domain, that effectively performs resizing in a differentiable way. Experiments on audio and image classification show the generality and effectiveness of our solution: we use DiffStride as a drop-in replacement to standard downsampling layers and outperform them. In particular, we show that introducing our layer into a ResNet-18 architecture allows keeping consistent high performance on CIFAR10, CIFAR100 and ImageNet even when training starts from poor random stride configurations. Moreover, formulating strides as learnable variables allows us to introduce a regularization term that controls the computational complexity of the architecture. We show how this regularization allows trading off accuracy for efficiency on ImageNet.

****

Nested Policy Reinforcement Learning for Clinical Decision Support

Aishwarya Mandyam,Andrew Jones,Krzysztof Laudanski,Barbara Engelhardt

Off-policy reinforcement learning (RL) has proven to be a powerful framework for guiding agents' actions in environments with stochastic rewards and unknown or noisy state dynamics. In many real-world settings, these agents must operate in multiple environments, each with slightly different dynamics. For example, we may be interested in developing policies to guide medical treatment for patients with and without a given disease, or policies to navigate curriculum design for students with and without a learning disability. Here, we introduce nested policy fitted Q-iteration (NFQI), an RL framework that finds optimal policies in environments that exhibit such a structure. Our approach develops a nested Q-value function that takes advantage of the shared structure between two groups of observations from two separate environments while allowing their policies to be distinct from one another. We find that NFQI yields policies that rely on relevant features and perform at least as well as a policy that does not consider group structure. We demonstrate NFQI's performance using an OpenAI Gym environment and a clinical decision making RL task. Our results suggest that NFQI can develop policies that are better suited to many real-world clinical environments.

****

On Lottery Tickets and Minimal Task Representations in Deep Reinforcement Learning

Marc Vischer,Robert Tjarko Lange,Henning Sprekeler

The lottery ticket hypothesis questions the role of overparameterization in supervised deep learning. But how is the performance of winning lottery tickets affected by the distributional shift inherent to reinforcement learning problems? In this work, we address this question by comparing sparse agents who have to address the non-stationarity of the exploration-exploitation problem with supervised agents trained to imitate an expert. We show that feed-forward networks trained with behavioural cloning compared to reinforcement learning can be pruned to higher levels of sparsity without performance degradation. This suggests that in order to solve the RL-specific distributional shift agents require more degrees of freedom. Using a set of carefully designed baseline conditions, we find that the majority of the lottery ticket effect in both learning paradigms can be attributed to the identified mask rather than the weight initialization. The input layer mask selectively prunes entire input dimensions that turn out to be irrelevant for the task at hand. At a moderate level of sparsity the mask identified by iterative magnitude pruning yields minimal task-relevant representations, i.e., an interpretable inductive bias. Finally, we propose a simple initialization rescaling which promotes the robust identification of sparse task representations in low-dimensional control tasks.

****

Brittle interpretations: The Vulnerability of TCAV and Other Concept-based Expla

inability Tools to Adversarial Attack
Davis Brown,Henry Kvinge

Methods for model explainability have become increasingly critical for testing the fairness and soundness of deep learning. A number of explainability techniques have been developed which use a set of examples to represent a human-interpretable concept in a model's activations. In this work we show that these explainability methods can suffer the same vulnerability to adversarial attacks as the models they are meant to analyze. We demonstrate this phenomenon on two well-known concept-based approaches to the explainability of deep learning models: TCAV and faceted feature visualization. We show that by carefully perturbing the examples of the concept that is being investigated, we can radically change the output of the interpretability method, e.g. showing that stripes are not an important factor in identifying images of a zebra. Our work highlights the fact that in safety-critical applications, there is need for security around not only the machine learning pipeline but also the model interpretation process.

****

Partial Information as Full: Reward Imputation with Sketching in Bandits
Xiao Zhang,Ninglu Shao,Zihua Si,Jun Xu,Wenhan Wang,hanjing su,Ji-Rong Wen

We focus on the setting of contextual batched bandit (CBB), where a batch of rewards is observed from the environment in each episode. But these rewards are partial-information feedbacks where the rewards of the non-executed actions are unobserved. Existing approaches for CBB usually ignore the potential rewards of the non-executed actions, resulting in feedback information being underutilized. In this paper, we propose an efficient reward imputation approach using sketching in CBB, which completes the unobserved rewards with the imputed rewards approximating the full-information feedbacks. Specifically, we formulate the reward imputation as a problem of imputation regularized ridge regression, which captures the feedback mechanisms of both the non-executed and executed actions. To reduce the time complexity of reward imputation on a large batch of data, we use randomized sketching for solving the regression problem of imputation. We prove that the proposed reward imputation approach obtains a relative-error bound for sketching approximation, achieves an instantaneous regret with an exponentially-decaying bias and a smaller variance than that without reward imputation, and enjoys a sublinear regret bound against the optimal policy. Moreover, we present two extensions of our approach, including the rate-scheduled version and the version for nonlinear rewards, which makes our approach more feasible. Experimental results demonstrated that our approach can outperform the state-of-the-art baselines on a synthetic dataset, the Criteo dataset, and a dataset from a commercial app.

****

Source-Target Unified Knowledge Distillation for Memory-Efficient Federated Domain Adaptation on Edge Devices
Xiaochen Zhou,Yuchuan Tian,Xudong Wang

To support local inference on an edge device, it is necessary to deploy a compact machine learning model on such a device.

When such a compact model is applied to a new environment, its inference accuracy can be degraded if the target data from the new environment have a different distribution from the source data that are used for model training.

To ensure high inference accuracy in the new environment, it is indispensable to adapt the compact model to the target data.

However, to protect users' privacy, the target data cannot be sent to a centralized server for joint training with the source data. Furthermore, utilizing the target data to directly train the compact model cannot achieve sufficient inference accuracy due to its low model capacity.

To this end, a scheme called source-target unified knowledge distillation (STU-KD) is developed in this paper. It starts with a large pretrained model loaded onto the edge device, and then the knowledge of the large model is transferred to the compact model via knowledge distillation.

Since training the large model leads to large memory consumption, a domain adaptation method called lite-residual hypothesis transfer is designed to achieve memory-efficient adaptation to the target data on the edge device. Moreover, to pre vent the compact model from forgetting the knowledge of the source data during knowledge distillation, a collaborative knowledge distillation (Co-KD) method is developed to unify the source data on the server and the target data on the edge device to train the compact model. STU-KD can be easily integrated with secure aggregation so that the server cannot obtain the true model parameters of the compact model. Extensive experiments conducted upon several objective recognition tasks show that STU-KD can improve the inference accuracy by up to 14.7%, as compared to the state-of-the-art schemes. Results also reveal that the inference accuracy of the compact model is not impacted by incorporating secure aggregation into STU-KD.

****

DriPP: Driven Point Processes to Model Stimuli Induced Patterns in M/EEG Signals

Cédric Allain,Alexandre Gramfort,Thomas Moreau

The quantitative analysis of non-invasive electrophysiology signals from electroencephalography (EEG) and magnetoencephalography (MEG) boils down to the identification of temporal patterns such as evoked responses, transient bursts of neural oscillations but also blinks or heartbeats for data cleaning. Several works have shown that these patterns can be extracted efficiently in an unsupervised way, e.g., using Convolutional Dictionary Learning. This leads to an event-based description of the data. Given these events, a natural question is to estimate how their occurrences are modulated by certain cognitive tasks and experimental manipulations. To address it, we propose a point process approach. While point processes have been used in neuroscience in the past, in particular for single cell recordings (spike trains), techniques such as Convolutional Dictionary Learning make them amenable to human studies based on EEG/MEG signals. We develop a novel statistical point process model – called driven temporal point processes (DriPP) – where the intensity function of the point process model is linked to a set of point processes corresponding to stimulation events. We derive a fast and principled expectation-maximization algorithm to estimate the parameters of this model. Simulations reveal that model parameters can be identified from long enough signals. Results on standard MEG datasets demonstrate that our methodology reveals event-related neural responses – both evoked and induced – and isolates non-task specific temporal patterns.

****

State-Action Joint Regularized Implicit Policy for Offline Reinforcement Learning

Shentao Yang,Zhendong Wang,Huangjie Zheng,Mingyuan Zhou

Offline reinforcement learning enables learning from a fixed dataset, without further interactions with the environment. The lack of environmental interactions makes the policy training vulnerable to state-action pairs far from the training dataset and prone to missing rewarding actions. For training more effective agents, we propose a framework that supports learning a flexible and well-regularized policy, which consists of a fully implicit policy and a regularization through the state-action visitation frequency induced by the current policy and that induced by the data-collecting behavior policy. We theoretically show the equivalence between policy-matching and state-action-visitation matching, and thus the compatibility of many prior work with our framework. An effective instantiation of our framework through the GAN structure is provided, together with some techniques to explicitly smooth the state-action mapping for robust generalization beyond the static dataset. Extensive experiments and ablation study on the D4RL dataset validate our framework and the effectiveness of our algorithmic designs.

****

Stiffness-aware neural network for learning Hamiltonian systems

SENWEI Liang,Zhongzhan Huang,Hong Zhang

We propose stiffness-aware neural network (SANN), a new method for learning Hamiltonian dynamical systems from data. SANN identifies and splits the training data into stiff and nonstiff portions based on a stiffness-aware index, a simple, yet effective metric we introduce to quantify the stiffness of the dynamical system. This classification along with a resampling technique allows us to apply different time integration strategies such as step size adaptation to better captur e the dynamical characteristics of the Hamiltonian vector fields. We evaluate SANN on complex physical systems including a three-body problem and billiard model. We show that SANN is more stable and can better preserve energy when compared with the state-of-the-art methods, leading to significant improvement in accuracy.

****

Learning Representation for Bayesian Optimization with Collision-free Regularization
Fengxue Zhang,Brian Nord,Yuxin Chen
Bayesian Optimization has been challenged by the large-scale and high-dimensional datasets, which are common in real-world scenarios. Recent works attempt to handle such input by applying neural networks ahead of the classical Gaussian process to learn a (low-dimensional) latent representation. We show that even with proper network design, such learned representation often leads to collision in the latent space: two points with significantly different observations collide in the learned latent space, leading to degraded optimization performance. To address this issue, we propose LOCo, an efficient deep Bayesian optimization framework which employs a novel regularizer to reduce the collision in the learned latent space and encourage the mapping from the latent space to the objective value to be Lipschitz continuous. LOCo takes in pairs of data points and penalizes those too close in the latent space compared to their target space distance. We provide a rigorous theoretical justification for LOCo by inspecting the regret of this dynamic-embedding-based Bayesian optimization algorithm, where the neural network is iteratively retrained with the regularizer. Our empirical results further demonstrate the effectiveness of LOCo on several synthetic and real-world benchmark Bayesian optimization tasks.

****

Adversarial Attacks on Spiking Convolutional Networks for Event-based Vision
Julian Büchel,Gregor Lenz,Yalun Hu,Sadique Sheik,Martino Sorbaro
Event-based sensing using dynamic vision sensors is gaining traction in low-power vision applications. Spiking neural networks work well with the sparse nature of event-based data and suit deployment on low-power neuromorphic hardware. Being a nascent field, the sensitivity of spiking neural networks to potentially malicious adversarial attacks has received very little attention so far. In this work, we show how white-box adversarial attack algorithms can be adapted to the discrete and sparse nature of event-based visual data, and to the continuous-time setting of spiking neural networks. We test our methods on the N-MNIST and IBM Gestures neuromorphic vision datasets and show adversarial perturbations achieve a high success rate, while injecting a relatively small number of appropriately placed events. We also verify, for the first time, the effectiveness of these perturbations directly on neuromorphic hardware. Finally, we discuss the properties of the resulting perturbations and possible future directions.

****

Self-supervised Learning is More Robust to Dataset Imbalance
Hong Liu,Jeff Z. HaoChen,Adrien Gaidon,Tengyu Ma
Self-supervised learning (SSL) is a scalable way to learn general visual representations since it learns without labels. However, large-scale unlabeled datasets in the wild often have long-tailed label distributions, where we know little about the behavior of SSL. In this work, we systematically investigate self-supervised learning under dataset imbalance. First, we find via extensive experiments that off-the-shelf self-supervised representations are already more robust to class imbalance than supervised representations. The performance gap between balanced and imbalanced pre-training with SSL is significantly smaller than the gap with supervised learning, across sample sizes, for both in-domain and, especially, out-of-domain evaluation. Second, towards understanding the robustness of SSL, we hypothesize that SSL learns richer features from frequent data: it may learn label-irrelevant-but-transferable features that help classify the rare classes and downstream tasks. In contrast, supervised learning has no incentive to learn features irrelevant to the labels from frequent examples. We validate this hypothesis with semi-synthetic experiments as well as rigorous mathematical analyses

on a simplified setting. Third, inspired by the theoretical insights, we devise a re-weighted regularization technique that consistently improves the SSL representation quality on imbalanced datasets with several evaluation criteria, closing the small gap between balanced and imbalanced datasets with the same number of examples.

****

CoST: Contrastive Learning of Disentangled Seasonal-Trend Representations for Time Series Forecasting

Gerald Woo,Chenghao Liu,Doyen Sahoo,Akshat Kumar,Steven Hoi

Deep learning has been actively studied for time series forecasting, and the mainstream paradigm is based on the end-to-end training of neural network architectures, ranging from classical LSTM/RNNs to more recent TCNs and Transformers. Motivated by the recent success of representation learning in computer vision and natural language processing, we argue that a more promising paradigm for time series forecasting, is to first learn disentangled feature representations, followed by a simple regression fine-tuning step -- we justify such a paradigm from a causal perspective. Following this principle, we propose a new time series representation learning framework for long sequence time series forecasting named CoST, which applies contrastive learning methods to learn disentangled seasonal-trend representations. CoST comprises both time domain and frequency domain contrastive losses to learn discriminative trend and seasonal representations, respectively. Extensive experiments on real-world datasets show that CoST consistently outperforms the state-of-the-art methods by a considerable margin, achieving a 21.3% improvement in MSE on multivariate benchmarks. It is also robust to various choices of backbone encoders, as well as downstream regressors. Code is available at https://github.com/salesforce/CoST.

****

CoordX: Accelerating Implicit Neural Representation with a Split MLP Architecture

Ruofan Liang,Hongyi Sun,Nandita Vijaykumar

Implicit neural representations with multi-layer perceptrons (MLPs) have recently gained prominence for a wide variety of tasks such as novel view synthesis and 3D object representation and rendering. However, a significant challenge with these representations is that both training and inference with an MLP over a large number of input coordinates to learn and represent an image, video, or 3D object, require large amounts of computation and incur long processing times. In this work, we aim to accelerate inference and training of coordinate-based MLPs for implicit neural representations by proposing a new split MLP architecture, CoordX. With CoordX, the initial layers are split to learn each dimension of the input coordinates separately. The intermediate features are then fused by the last layers to generate the learned signal at the corresponding coordinate point. This significantly reduces the amount of computation required and leads to large speedups in training and inference, while achieving similar accuracy as the baseline MLP. This approach thus aims at first learning functions that are a decomposition of the original signal and then fusing them to generate the learned signal. Our proposed architecture can be generally used for many implicit neural representation tasks with no additional memory overheads. We demonstrate a speedup of up to 2.92x compared to the baseline model for image, video, and 3D shape representation and rendering tasks.

****

EBM Life Cycle: MCMC Strategies for Synthesis, Defense, and Density Modeling

Mitch Hill,Jonathan Craig Mitchell,Chu Chen,Yuan Du,Mubarak Shah,Song-Chun Zhu

This work presents strategies to learn an Energy-Based Model (EBM) according to the desired length of its MCMC sampling trajectories. MCMC trajectories of different lengths correspond to models with different purposes. Our experiments cover three different trajectory magnitudes and learning outcomes: 1) shortrun sampling for image generation; 2) midrun sampling for classifier-agnostic adversarial defense; and 3) longrun sampling for principled modeling of image probability densities. To achieve these outcomes, we introduce three novel methods of MCMC initialization for negative samples used in Maximum Likelihood (ML) learning. With

standard network architectures and an unaltered ML objective, our MCMC initialization methods alone enable significant performance gains across the three applications that we investigate. Our results include state-of-the-art FID scores for unnormalized image densities on the CIFAR-10 and ImageNet datasets; state-of-the-art adversarial defense on CIFAR-10 among purification methods and the first EBM defense on ImageNet; and scalable techniques for learning valid probability densities.
****

Intriguing Properties of Input-dependent Randomized Smoothing
Peter Súkeník,Aleksei Kuvshinov,Stephan Günnemann

Randomized smoothing is currently considered the state-of-the-art method to obtain certifiably robust classifiers. Despite its remarkable performance, the method is associated with various serious problems such as ``certified accuracy waterfalls'', certification vs.\ accuracy trade-off, or even fairness issues. Input-dependent smoothing approaches have been proposed to overcome these flaws. However, we demonstrate that these methods lack formal guarantees and so the resulting certificates are not justified. We show that the input-dependent smoothing, in general, suffers from the curse of dimensionality, forcing the variance function to have low semi-elasticity. On the other hand, we provide a theoretical and practical framework that enables the usage of input-dependent smoothing even in the presence of the curse of dimensionality, under strict restrictions. We present one concrete design of the smoothing variance and test it on CIFAR10 and MNIST. Our design solves some of the problems of classical smoothing and is formally underlined, yet further improvement of the design is still necessary.
****

Auditing AI models for Verified Deployment under Semantic Specifications
Homanga Bharadhwaj,De-An Huang,Chaowei Xiao,Anima Anandkumar,Animesh Garg

Auditing trained deep learning (DL) models prior to deployment is vital for preventing unintended consequences. One of the biggest challenges in auditing is in understanding how we can obtain human-interpretable specifications that are directly useful to the end-user. We address this challenge through a sequence of semantically-aligned unit tests, where each unit test verifies whether a predefined specification (e.g., accuracy over 95\%) is satisfied with respect to controlled and semantically aligned variations in the input space (e.g., in face recognition, the angle relative to the camera). We perform these unit tests by directly verifying the semantically aligned variations in an interpretable latent space of a generative model by building a bridge with the DL model. Our framework, AuditAI, bridges the gap between interpretable formal verification and scalability. With evaluations on four different datasets, covering images of chest X-rays, human faces, ImageNet classes, and towers, we show how AuditAI allows us to obtain controlled variations for verification and certified training. We address the limitations of the standard approach of verifying using only pixel-space perturbations.
****

Plant 'n' Seek: Can You Find the Winning Ticket?
Jonas Fischer,Rebekka Burkholz

The lottery ticket hypothesis has sparked the rapid development of pruning algorithms that aim to reduce the computational costs associated with deep learning during training and model deployment. Currently, such algorithms are primarily evaluated on imaging data, for which we lack ground truth information and thus the understanding of how sparse lottery tickets could be. To fill this gap, we develop a framework that allows us to plant and hide winning tickets with desirable properties in randomly initialized neural networks. To analyze the ability of state-of-the-art pruning to identify tickets of extreme sparsity, we design and hide such tickets solving four challenging tasks. In extensive experiments, we observe similar trends as in imaging studies, indicating that our framework can provide transferable insights into realistic problems. Additionally, we can now see beyond such relative trends and highlight limitations of current pruning methods. Based on our results, we conclude that the current limitations in ticket sparsity are likely of algorithmic rather than fundamental nature. We anticipate tha

t comparisons to planted tickets will facilitate future developments of efficient pruning algorithms.
****

Coherence-based Label Propagation over Time Series for Accelerated Active Learning

Yooju Shin,Susik Yoon,Sundong Kim,Hwanjun Song,Jae-Gil Lee,Byung Suk Lee

Time-series data are ubiquitous these days, but lack of the labels in time-series data is regarded as a hurdle for its broad applicability. Meanwhile, active learning has been successfully adopted to reduce the labeling efforts in various tasks. Thus, this paper addresses an important issue, time-series active learning. Inspired by the temporal coherence in time-series data, where consecutive data points tend to have the same label, our label propagation framework, called TCLP, automatically assigns a queried label to the data points within an accurately estimated time-series segment, thereby significantly boosting the impact of an individual query. Compared with traditional time-series active learning, TCLP is shown to improve the classification accuracy by up to 7.1 times when only 0.8% of data points in the entire time series are queried for their labels.
****

A Class of Short-term Recurrence Anderson Mixing Methods and Their Applications

Fuchao Wei,Chenglong Bao,Yang Liu

Anderson mixing (AM) is a powerful acceleration method for fixed-point iterations, but its computation requires storing many historical iterations. The extra memory footprint can be prohibitive when solving high-dimensional problems in a resource-limited machine. To reduce the memory overhead, we propose a novel class of short-term recurrence AM methods (ST-AM). The ST-AM methods only store two previous iterations with cheap corrections. We prove that the basic version of ST-AM is equivalent to the full-memory AM in strongly convex quadratic optimization, and with minor changes it has local linear convergence for solving general nonlinear fixed-point problems. We further analyze the convergence properties of the regularized ST-AM for nonconvex (stochastic) optimization. Finally, we apply ST-AM to several applications including solving root-finding problems and training neural networks. Experimental results show that ST-AM is competitive with the long-memory AM and outperforms many existing optimizers.
****

The Geometry of Memoryless Stochastic Policy Optimization in Infinite-Horizon POMDPs

Johannes Müller,Guido Montufar

We consider the problem of finding the best memoryless stochastic policy for an infinite-horizon partially observable Markov decision process (POMDP) with finite state and action spaces with respect to either the discounted or mean reward criterion. We show that the (discounted) state-action frequencies and the expected cumulative reward are rational functions of the policy, whereby the degree is determined by the degree of partial observability. We then describe the optimization problem as a linear optimization problem in the space of feasible state-action frequencies subject to polynomial constraints that we characterize explicitly. This allows us to address the combinatorial and geometric complexity of the optimization problem using recent tools from polynomial optimization. In particular, we demonstrate how the partial observability constraints can lead to multiple smooth and non-smooth local optimizers and we estimate the number of critical points.
****

INFERNO: Inferring Object-Centric 3D Scene Representations without Supervision

Lluis Castrejon,Nicolas Ballas,Aaron Courville

We propose INFERNO, a method to infer object-centric representations of visual scenes without relying on annotations. Our method learns to decompose a scene into multiple objects, each object having a structured representation that disentangles its shape, appearance and 3D pose. To impose this structure we rely on recent advances in neural 3D rendering. Each object representation defines a localized neural radiance field that is used to generate 2D views of the scene through a differentiable rendering process. Our model is subsequently trained by minimiz ing a reconstruction loss between inputs and corresponding rendered scenes. We empirically show that INFERNO discovers objects in a scene without supervision. We also validate the interpretability of the learned representations by manipulating inferred scenes and showing the corresponding effect in the rendered output. Finally, we demonstrate the usefulness of our 3D object representations in a visual reasoning task using the CATER dataset.

****

Efficient Sharpness-aware Minimization for Improved Training of Neural Networks

Jiawei Du,Hanshu Yan,Jiashi Feng,Joey Tianyi Zhou,Liangli Zhen,Rick Siow Mong Goh,Vincent Tan

Overparametrized Deep Neural Networks (DNNs) often achieve astounding performances, but may potentially result in severe generalization error. Recently, the relation between the sharpness of the loss landscape and the generalization error has been established by Foret et al. (2020), in which the Sharpness Aware Minimizer (SAM) was proposed to mitigate the degradation of the generalization. Unfortunately, SAM’s computational cost is roughly double that of base optimizers, such as Stochastic Gradient Descent (SGD). This paper thus proposes Efficient Sharpness Aware Minimizer (ESAM), which boosts SAM’s efficiency at no cost to its generalization performance. ESAM includes two novel and efficient training strategies—StochasticWeight Perturbation and Sharpness-Sensitive Data Selection. In the former, the sharpness measure is approximated by perturbing a stochastically chosen set of weights in each iteration; in the latter, the SAM loss is optimized using only a judiciously selected subset of data that is sensitive to the sharpness. We provide theoretical explanations as to why these strategies perform well. We also show, via extensive experiments on the CIFAR and ImageNet datasets, that ESAM enhances the efficiency over SAM from requiring 100% extra computations to 40% vis-`a-vis base optimizers, while test accuracies are preserved or even improved.

****

Lipschitz-constrained Unsupervised Skill Discovery

Seohong Park,Jongwook Choi,Jaekyeom Kim,Honglak Lee,Gunhee Kim

We study the problem of unsupervised skill discovery, whose goal is to learn a set of diverse and useful skills with no external reward. There have been a number of skill discovery methods based on maximizing the mutual information (MI) between skills and states. However, we point out that their MI objectives usually prefer static skills to dynamic ones, which may hinder the application for downstream tasks. To address this issue, we propose Lipschitz-constrained Skill Discovery (LSD), which encourages the agent to discover more diverse, dynamic, and far-reaching skills. Another benefit of LSD is that its learned representation function can be utilized for solving goal-following downstream tasks even in a zero-shot manner — i.e., without further training or complex planning. Through experiments on various MuJoCo robotic locomotion and manipulation environments, we demonstrate that LSD outperforms previous approaches in terms of skill diversity, state space coverage, and performance on seven downstream tasks including the challenging task of following multiple goals on Humanoid. Our code and videos are available at https://shpark.me/projects/lsd/.

****

Learning Generalizable Representations for Reinforcement Learning via Adaptive Meta-learner of Behavioral Similarities

Jianda Chen,Sinno Pan

How to learn an effective reinforcement learning-based model for control tasks from high-level visual observations is a practical and challenging problem. A key to solving this problem is to learn low-dimensional state representations from observations, from which an effective policy can be learned. In order to boost the learning of state encoding, recent works are focused on capturing behavioral similarities between state representations or applying data augmentation on visual observations. In this paper, we propose a novel meta-learner-based framework for representation learning regarding behavioral similarities for reinforcement learning. Specifically, our framework encodes the high-dimensional observations into two decomposed embeddings regarding reward and dynamics in a Markov Decisio n Process (MDP). A pair of meta-learners are developed, one of which quantifies the reward similarity and the other quantifies dynamics similarity over the correspondingly decomposed embeddings. The meta-learners are self-learned to update the state embeddings by approximating two disjoint terms in on-policy bisimulation metric. To incorporate the reward and dynamics terms, we further develop a strategy to adaptively balance their impacts based on different tasks or environments. We empirically demonstrate that our proposed framework outperforms state-of-the-art baselines on several benchmarks, including conventional DM Control Suite, Distracting DM Control Suite and a self-driving task CARLA.

****

Effective Model Sparsification by Scheduled Grow-and-Prune Methods

Xiaolong Ma,Minghai Qin,Fei Sun,Zejiang Hou,Kun Yuan,Yi Xu,Yanzhi Wang,Yen-Kuang Chen,Rong Jin,Yuan Xie

Deep neural networks (DNNs) are effective in solving many real-world problems. Larger DNN models usually exhibit better quality (e.g., accuracy) but their excessive computation results in long inference time. Model sparsification can reduce the computation and memory cost while maintaining model quality. Most existing sparsification algorithms unidirectionally remove weights, while others randomly or greedily explore a small subset of weights in each layer for pruning. The limitations of these algorithms reduce the level of achievable sparsity. In addition, many algorithms still require pre-trained dense models and thus suffer from large memory footprint. In this paper, we propose a novel scheduled grow-and-prune (GaP) methodology without having to pre-train a dense model. It addresses the shortcomings of the previous works by repeatedly growing a subset of layers to dense and then pruning them back to sparse after some training. Experiments show that the models pruned using the proposed methods match or beat the quality of the highly optimized dense models at 80% sparsity on a variety of tasks, such as image classification, objective detection, 3D object part segmentation, and translation. They also outperform other state-of-the-art (SOTA) methods for model sparsification. As an example, a 90% non-uniform sparse ResNet-50 model obtained via GaP achieves 77.9% top-1 accuracy on ImageNet, improving the previous SOTA results by 1.5%. Code available at: https://github.com/boone891214/GaP.

****

Assessing two novel distance-based loss functions for few-shot image classification

Mauricio Mendez Ruiz,Gilberto Ochoa Ruiz,Andres Mendez Vazquez,Jorge Gonzalez-Zapata

Few-shot learning is a challenging area of research which aims to learn new concepts with only a few labeled samples of data. Recent works based on metric-learning approaches benefit from the meta-learning process in which we have episodic tasks conformed by support set (training) and query set (test), and the objective is to learn a similarity comparison metric between those sets. Due to the lack of data, the learning process of the embedding network becomes an important part of the few-shot task. In this work, we propose two different loss functions which consider the importance of the embedding vectors by looking at the intra-class and inter-class distance between the few data. The first loss function is the Proto-Triplet Loss, which is based on the original triplet loss with the modifications needed to better work on few-shot scenarios. The second loss function is based on an inter and intra class nearest neighbors score, which help us to know the quality of embeddings obtained from the trained network. Extensive experimental results on the miniImagenNet benchmark increase the accuracy performance from other metric-based few-shot learning methods by a margin of 2%, demonstrating the capability of these loss functions to allow the network to generalize better to previously unseen classes.

****

FILIP: Fine-grained Interactive Language-Image Pre-Training

Lewei Yao,Runhui Huang,Lu Hou,Guansong Lu,Minzhe Niu,Hang Xu,Xiaodan Liang,Zhenguo Li,Xin Jiang,Chunjing Xu

Unsupervised large-scale vision-language pre-training has shown promising advances on various downstream tasks. Existing methods often model the cross-modal int eraction either via the similarity of the global feature of each modality which misses sufficient information, or finer-grained interactions using cross/self-attention upon visual and textual tokens. However, cross/self-attention suffers from inferior efficiency in both training and inference. In this paper, we introduce a large-scale Fine-grained Interactive Language-Image Pre-training (FILIP) to achieve finer-level alignment through a cross-modal late interaction mechanism, which uses a token-wise maximum similarity between visual and textual tokens to guide the contrastive objective. FILIP successfully leverages the finer-grained expressiveness between image patches and textual words by modifying only contrastive loss, while simultaneously gaining the ability to pre-compute image and text representations offline at inference, keeping both large-scale training and inference efficient. Furthermore, we construct a new large-scale image-text pair dataset called FILIP300M for pre-training. Experiments show that FILIP achieves state-of-the-art performance on multiple downstream vision-language tasks including zero-shot image classification and image-text retrieval. The visualization on word-patch alignment further shows that FILIP can learn meaningful fine-grained features with promising localization ability.

****

Implicit Bias of Linear Equivariant Networks

Hannah Lawrence,Kristian Georgiev,Andrew Dienes,Bobak Kiani

Group equivariant convolutional neural networks (G-CNNs) are generalizations of convolutional neural networks (CNNs) which excel in a wide range of scientific and technical applications by explicitly encoding particular group symmetries, such as rotations and permutations, in their architectures. Although the success of G-CNNs is driven by the explicit symmetry bias of their convolutional architecture, a recent line of work has proposed that the implicit bias of training algorithms on a particular parameterization (or architecture) is key to understanding generalization for overparameterized neural nets. In this context, we show that L-layer full-width linear G-CNNs trained via gradient descent in a binary classification task converge to solutions with low-rank Fourier matrix coefficients, regularized by the $2/L$-Schatten matrix norm. Our work strictly generalizes previous analysis on the implicit bias of linear CNNs to linear G-CNNs over all finite groups, including the challenging setting of non-commutative symmetry groups (such as permutations). We validate our theorems via experiments on a variety of groups and empirically explore more realistic nonlinear networks, which locally capture similar regularization patterns. Finally, we provide intuitive interpretations of our Fourier-space implicit regularization results in real space via uncertainty principles.

****

Information Prioritization through Empowerment in Visual Model-based RL

Homanga Bharadhwaj,Mohammad Babaeizadeh,Dumitru Erhan,Sergey Levine

Model-based reinforcement learning (RL) algorithms designed for handling complex visual observations typically learn some sort of latent state representation, either explicitly or implicitly. Standard methods of this sort do not distinguish between functionally relevant aspects of the state and irrelevant distractors, instead aiming to represent all available information equally. We propose a modified objective for model-based RL that, in combination with mutual information maximization, allows us to learn representations and dynamics for visual model-based RL without reconstruction in a way that explicitly prioritizes functionally relevant factors. The key principle behind our design is to integrate a term inspired by variational empowerment into a state-space learning model based on mutual information. This term prioritizes information that is correlated with action, thus ensuring that functionally relevant factors are captured first. Furthermore, the same empowerment term also promotes faster exploration during the RL process, especially for sparse-reward tasks where the reward signal is insufficient to drive exploration in the early stages of learning. We evaluate the approach on a suite of vision-based robot control tasks with natural video backgrounds, and show that the proposed prioritized information objective outperforms state-of-the-art model based RL approaches by an average of 20\% in terms of episodic returns at 1M environment interactions with 30\% higher sample efficiency at 100k interactions.

****

Efficient Active Search for Combinatorial Optimization Problems

André Hottung,Yeong-Dae Kwon,Kevin Tierney

Recently numerous machine learning based methods for combinatorial optimization problems have been proposed that learn to construct solutions in a sequential decision process via reinforcement learning. While these methods can be easily combined with search strategies like sampling and beam search, it is not straightforward to integrate them into a high-level search procedure offering strong search guidance. Bello et al. (2016) propose active search, which adjusts the weights of a (trained) model with respect to a single instance at test time using reinforcement learning. While active search is simple to implement, it is not competitive with state-of-the-art methods because adjusting all model weights for each test instance is very time and memory intensive. Instead of updating all model weights, we propose and evaluate three efficient active search strategies that only update a subset of parameters during the search. The proposed methods offer a simple way to significantly improve the search performance of a given model and outperform state-of-the-art machine learning based methods on combinatorial problems, even surpassing the well-known heuristic solver LKH3 on the capacitated vehicle routing problem. Finally, we show that (efficient) active search enables learned models to effectively solve instances that are much larger than those seen during training.

****

Ancestral protein sequence reconstruction using a tree-structured Ornstein-Uhlenbeck variational autoencoder

Lys Sanz Moreta,Ola Rønning,Ahmad Salim Al-Sibahi,Jotun Hein,Douglas Theobald,Thomas Hamelryck

We introduce a deep generative model for representation learning of biological sequences that, unlike existing models, explicitly represents the evolutionary process. The model makes use of a tree-structured Ornstein-Uhlenbeck process, obtained from a given phylogenetic tree, as an informative prior for a variational autoencoder. We show the model performs well on the task of ancestral sequence reconstruction of single protein families. Our results and ablation studies indicate that the explicit representation of evolution using a suitable tree-structured prior has the potential to improve representation learning of biological sequences considerably. Finally, we briefly discuss extensions of the model to genomic-scale data sets and the case of a latent phylogenetic tree.

****

Imbalanced Adversarial Training with Reweighting

Wentao Wang,Han Xu,Xiaorui Liu,Yaxin Li,Bhavani Thuraisingham,Jiliang Tang

Adversarial training has been empirically proven to be one of the most effective and reliable defense methods against adversarial attacks. However, the majority of existing studies are focused on balanced datasets, where each class has a similar amount of training examples. Research on adversarial training with imbalanced training datasets is rather limited. As the initial effort to investigate this problem, we reveal the facts that adversarially trained models present two distinguished behaviors from naturally trained models in imbalanced datasets: (1) Compared to natural training, adversarially trained models can suffer much worse performance on under-represented classes, when the training dataset is extremely imbalanced. (2) Traditional reweighting strategies may lose efficacy to deal with the imbalance issue for adversarial training. For example, upweighting under-represented classes will drastically hurt the model’s performance on well-represented classes, and as a result, finding an optimal reweighting value can be tremendously challenging. In this paper, to further understand our observations, we theoretically show that the poor data separability is one key reason causing this strong tension between under-represented and well-represented classes. Motivated by this finding, we propose Separable Reweighted Adversarial Training (SRAT) to facilitate adversarial training under imbalanced scenarios, by learning more separable features for different classes. Extensive experiments on various datasets verify the effectiveness of the proposed framework.

****

Stabilized Self-training with Negative Sampling on Few-labeled Graph Data

Ziang Zhou,Jieming Shi,Shengzhong Zhang,Zengfeng Huang,Qing Li

Graph neural networks (GNNs) are designed for semi-supervised node classification on graphs where only a small subset of nodes have class labels. However, under extreme cases when very few labels are available (e.g., 1 labeled node per class), GNNs suffer from severe result quality degradation.

Specifically, we observe that existing GNNs suffer from unstable training process on few-labeled graph data, resulting to inferior performance on node classification. Therefore, we propose an effective framework, Stabilized self-training with Negative sampling (SN), which is applicable to existing GNNs to stabilize the training process and enhance the training data, and consequently, boost classification accuracy on graphs with few labeled data. In experiments, we apply our SN framework to two existing GNN base models (GCN and DAGNN) to get SNGCN and SNDAGNN, and evaluate the two methods against 13 existing solutions over 4 benchmarking datasets. Extensive experiments show that the proposed SN framework is highly effective compared with existing solutions, especially under settings with very few labeled data. In particular, on a benchmark dataset Cora with only 1 labeled node per class, while GCN only has 44.6% accuracy, SNGCN achieves 62.5% accuracy, improving GCN by 17.9%; SNDAGNN has accuracy 66.4%, improving that of the base model DAGNN (59.8%) by 6.6%.

****

Ab-Initio Potential Energy Surfaces by Pairing GNNs with Neural Wave Functions

Nicholas Gao,Stephan Günnemann

Solving the Schrödinger equation is key to many quantum mechanical properties. However, an analytical solution is only tractable for single-electron systems. Recently, neural networks succeeded at modelling wave functions of many-electron systems. Together with the variational Monte-Carlo (VMC) framework, this led to solutions on par with the best known classical methods. Still, these neural methods require tremendous amounts of computational resources as one has to train a separate model for each molecular geometry. In this work, we combine a Graph Neural Network (GNN) with a neural wave function to simultaneously solve the Schrödinger equation for multiple geometries via VMC. This enables us to model continuous subsets of the potential energy surface with a single training pass. Compared to existing state-of-the-art networks, our Potential Energy Surface Network (PESNet) speeds up training for multiple geometries by up to 40 times while matching or surpassing their accuracy. This may open the path to accurate and orders of magnitude cheaper quantum mechanical calculations.

****

Self-GenomeNet: Self-supervised Learning with Reverse-Complement Context Prediction for Nucleotide-level Genomics Data

Hüseyin Anil Gündüz,Martin Binder,Xiao-Yin To,René Mreches,Philipp C. Münch,Alice C McHardy,Bernd Bischl,Mina Rezaei

We introduce Self-GenomeNet, a novel contrastive self-supervised learning method for nucleotide-level genomic data, which substantially improves the quality of the learned representations and performance compared to the current state-of-the-art deep learning frameworks. To the best of our knowledge, Self-GenomeNet is the first self-supervised framework that learns a representation of nucleotide-level genome data, using domain-specific characteristics. Our proposed method learns and parametrizes the latent space by leveraging the reverse-complement of genomic sequences. During the training procedure, we force our framework to capture semantic representations with a novel context network on top of intermediate features extracted by an encoder network. The network is trained with an unsupervised contrastive loss. Extensive experiments show that our method with self-supervised and semi-supervised settings is able to considerably outperform previous deep learning methods on different datasets and a public bioinformatics benchmark. Moreover, the learned representations generalize well when transferred to new datasets and tasks. The source code of the method and all the experiments are available at supplementary.

****

Learning to Solve Combinatorial Problems via Efficient Exploration

Thomas D Barrett,Christopher William Falke Parsonson,Alexandre Laterre

From logistics to the natural sciences, combinatorial optimisation on graphs underpins numerous real-world applications. Reinforcement learning (RL) has shown particular promise in this setting as it can adapt to specific problem structures and does not require pre-solved instances for these, often NP-hard, problems. However, state-of-the-art (SOTA) approaches typically suffer from severe scalability issues, primarily due to their reliance on expensive graph neural networks (GNNs) at each decision step. We introduce ECORD; a novel RL algorithm that alleviates this expense by restricting the GNN to a single pre-processing step, before entering a fast-acting exploratory phase directed by a recurrent unit. Experimentally, we demonstrate that ECORD achieves a new SOTA for RL algorithms on the Maximum Cut problem, whilst also providing orders of magnitude improvement in speed and scalability. Compared to the nearest competitor, ECORD reduces the optimality gap by up to 73% on 500 vertex graphs with a decreased wall-clock time. Moreover, ECORD retains strong performance when generalising to larger graphs with up to 10000 vertices.

****

3D Pre-training improves GNNs for Molecular Property Prediction

Hannes Stärk,Dominique Beaini,Gabriele Corso,Prudencio Tossou,Christian Dallago,Stephan Günnemann,Pietro Lio

Molecular property prediction is one of the fastest-growing applications of deep learning with critical real-world impacts. Including 3D molecular structure as input to learned models their performance for many molecular tasks. However, this information is infeasible to compute at the scale required by several real-world applications. We propose pre-training a model to reason about the geometry of molecules given only their 2D molecular graphs. Using methods from self-supervised learning, we maximize the mutual information between 3D summary vectors and the representations of a Graph Neural Network (GNN) such that they contain latent 3D information. During fine-tuning on molecules with unknown geometry, the GNN still generates implicit 3D information and can use it to improve downstream tasks. We show that 3D pre-training provides significant improvements for a wide range of properties, such as a 22% average MAE reduction on eight quantum mechanical properties. Moreover, the learned representations can be effectively transferred between datasets in different molecular spaces.

****

Training Structured Neural Networks Through Manifold Identification and Variance Reduction

Zih-Syuan Huang,Ching-pei Lee

This paper proposes an algorithm, RMDA, for training neural networks (NNs) with a regularization term for promoting desired structures. RMDA does not incur computation additional to proximal SGD with momentum, and achieves variance reduction without requiring the objective function to be of the finite-sum form. Through the tool of manifold identification from nonlinear optimization, we prove that after a finite number of iterations, all iterates of RMDA possess a desired structure identical to that induced by the regularizer at the stationary point of asymptotic convergence, even in the presence of engineering tricks like data augmentation that complicate the training process. Experiments on training NNs with structured sparsity confirm that variance reduction is necessary for such an identification, and show that RMDA thus significantly outperforms existing methods for this task. For unstructured sparsity, RMDA also outperforms a state-of-the-art pruning method, validating the benefits of training structured NNs through regularization.

Implementation of RMDA is available at https://www.github.com/zihsyuan1214/rmda.

****

The Neural Data Router: Adaptive Control Flow in Transformers Improves Systematic Generalization

Róbert Csordás,Kazuki Irie,Jürgen Schmidhuber

Despite progress across a broad range of applications, Transformers have limited success in systematic generalization. The situation is especially frustrating i n the case of algorithmic tasks, where they often fail to find intuitive solutions that route relevant information to the right node/operation at the right time in the grid represented by Transformer columns. To facilitate the learning of useful control flow, we propose two modifications to the Transformer architecture, copy gate and geometric attention. Our novel Neural Data Router (NDR) achieves 100% length generalization accuracy on the classic compositional table lookup task, as well as near-perfect accuracy on the simple arithmetic task and a new variant of ListOps testing for generalization across computational depths. NDR’s attention and gating patterns tend to be interpretable as an intuitive form of neural routing

****

Learning Explicit Credit Assignment for Multi-agent Joint Q-learning

Hangyu Mao,Jianye HAO,Dong Li,Jun Wang,Weixun Wang,Xiaotian Hao,Bin Wang,Kun Shao,Zhen Xiao,Wulong Liu

Multi-agent joint Q-learning based on Centralized Training with Decentralized Execution (CTDE) has become an effective technique for multi-agent cooperation. During centralized training, these methods are essentially addressing the multi-agent credit assignment problem. However, most of the existing methods \emph{implicitly} learn the credit assignment just by ensuring that the joint Q-value satisfies the Bellman optimality equation. In contrast, we formulate an \emph{explicit} credit assignment problem where each agent gives its suggestion about how to weight individual Q-values to explicitly maximize the joint Q-value, besides guaranteeing the Bellman optimality of the joint Q-value. In this way, we can conduct credit assignment among multiple agents and along the time horizon. Theoretically, we give a gradient ascent solution for this problem. Empirically, we instantiate the core idea with deep neural networks and propose Explicit Credit Assignment joint Q-learning (ECAQ) to facilitate multi-agent cooperation in complex problems. Extensive experiments justify that ECAQ achieves interpretable credit assignment and superior performance compared to several advanced baselines.

****

LEAN: graph-based pruning for convolutional neural networks by extracting longest chains

Richard Arnoud Schoonhoven,Allard Hendriksen,Daniel Pelt,Joost Batenburg

Neural network pruning techniques can substantially reduce the computational cost of applying convolutional neural networks (CNNs). Common pruning methods determine which convolutional filters to remove by ranking the filters individually, i.e., without taking into account their interdependence. In this paper, we advocate the viewpoint that pruning should consider the interdependence between series of consecutive operators. We propose the LongEst-chAiN (LEAN) method that prunes CNNs by using graph-based algorithms to select relevant chains of convolutions. A CNN is interpreted as a graph, with the operator norm of each operator as distance metric for the edges. LEAN pruning iteratively extracts the highest value path from the graph to keep. In our experiments, we test LEAN pruning on several image-to-image tasks, including the well-known CamVid dataset, and a real-world X-ray CT dataset. Results indicate that LEAN pruning can result in networks with similar accuracy but 3--20x fewer convolutional filters than networks pruned with methods that rank filters individually.

****

Automated Channel Pruning with Learned Importance

■ukasz Treszczotko,Pawel Kubik

Neural network pruning allows for significant reduction of model size and latency. However, most of the current network pruning methods do not consider channel interdependencies and a lot of manual adjustments are required before they can be applied to new network architectures. Moreover, these algorithms are often based on hand-picked, sometimes complicated heuristics and can require thousands of GPU computation hours. In this paper, we introduce a simple neural network pruning and fine-tuning framework that requires no manual heuristics, is highly efficient to train (2-6 times speed up compared to NAS-based competitors) and produces comparable performance. The framework contains 1) an automatic channel detection algorithm that groups the interdependent blocks of channels; 2) a non-itera

tive pruning algorithm that learns channel importance directly from feature maps while masking the coupled computational blocks using Gumbel-Softmax sampling and 3) a hierarchical knowledge distillation approach to fine-tune the pruned neural networks. We validate our pipeline on ImageNet classification, human segmentation and image denoising, creating lightweight and low latency models, easy to deploy on mobile devices. Using our pruning algorithm and hierarchical knowledge distillation for fine-tuning we are able to prune EfficientNet B0, EfficientNetV2 B0 and MobileNetV2 to 75% of their original FLOPs with no loss of accuracy on ImageNet. We release a set pruned backbones as Keras models - all of them proved beneficial when deployed in other projects.
****

On the Limitations of Multimodal VAEs

Imant Daunhawer,Thomas M. Sutter,Kieran Chin-Cheong,Emanuele Palumbo,Julia E Vogt

Multimodal variational autoencoders (VAEs) have shown promise as efficient generative models for weakly-supervised data. Yet, despite their advantage of weak supervision, they exhibit a gap in generative quality compared to unimodal VAEs, which are completely unsupervised. In an attempt to explain this gap, we uncover a fundamental limitation that applies to a large family of mixture-based multimodal VAEs. We prove that the sub-sampling of modalities enforces an undesirable upper bound on the multimodal ELBO and thereby limits the generative quality of the respective models. Empirically, we showcase the generative quality gap on both synthetic and real data and present the tradeoffs between different variants of multimodal VAEs. We find that none of the existing approaches fulfills all desired criteria of an effective multimodal generative model when applied on more complex datasets than those used in previous benchmarks. In summary, we identify, formalize, and validate fundamental limitations of VAE-based approaches for modeling weakly-supervised data and discuss implications for real-world applications.
****

Recursive Disentanglement Network

Yixuan Chen,Yubin Shi,Dongsheng Li,Yujiang Wang,Mingzhi Dong,Yingying Zhao,Robert Dick,Qin Lv,Fan Yang,Li Shang

Disentangled feature representation is essential for data-efficient learning. The feature space of deep models is inherently compositional. Existing β-VAE-based methods, which only apply disentanglement regularization to the resulting embedding space of deep models, cannot effectively regularize such compositional feature space, resulting in unsatisfactory disentangled results. In this paper, we formulate the compositional disentanglement learning problem from an information-theoretic perspective and propose a recursive disentanglement network (RecurD) that propagates regulatory inductive bias recursively across the compositional feature space during disentangled representation learning.

Experimental studies demonstrate that RecurD outperforms β-VAE and several of its state-of-the-art variants on disentangled representation learning and enables more data-efficient downstream machine learning tasks.
****

ADAVI: Automatic Dual Amortized Variational Inference Applied To Pyramidal Bayesian Models

Louis Rouillard,Demian Wassermann

Frequently, population studies feature pyramidally-organized data represented using Hierarchical Bayesian Models (HBM) enriched with plates. These models can become prohibitively large in settings such as neuroimaging, where a sample is composed of a functional MRI signal measured on 300 brain locations, across 4 measurement sessions, and 30 subjects, resulting in around 1 million latent parameters.

Such high dimensionality hampers the usage of modern, expressive flow-based techniques.

To infer parameter posterior distributions in this challenging class of problems

, we designed a novel methodology that automatically produces a variational family dual to a target HBM. This variational family, represented as a neural network, consists in the combination of an attention-based hierarchical encoder feeding summary statistics to a set of normalizing flows. Our automatically-derived neural network exploits exchangeability in the plate-enriched HBM and factorizes its parameter space. The resulting architecture reduces by orders of magnitude its parameterization with respect to that of a typical flow-based representation, while maintaining expressivity.

Our method performs inference on the specified HBM in an amortized setup: once trained, it can readily be applied to a new data sample to compute the parameters' full posterior.

We demonstrate the capability and scalability of our method on simulated data, as well as a challenging high-dimensional brain parcellation experiment. We also open up several questions that lie at the intersection between normalizing flows, SBI, structured Variational Inference, and inference amortization.

****

Inductive Lottery Ticket Learning for Graph Neural Networks

Yongduo Sui,Xiang Wang,Tianlong Chen,Xiangnan He,Tat-Seng Chua

Deep graph neural networks (GNNs) have gained increasing popularity, while usually suffer from unaffordable computations for real-world large-scale applications. Hence, pruning GNNs is of great need but largely unexplored. A recent work, UGS, studies lottery ticket learning for GNNs, aiming to find a subset of model parameters and graph structure that can best maintain the GNN performance. However, it is tailed for the transductive setting, failing to generalize to unseen graphs, which are common in inductive tasks like graph classification. In this work, we propose a simple and effective learning paradigm, Inductive Co-Pruning of GNNs (ICPG), to endow graph lottery tickets with inductive pruning capacity. To prune the input graphs, we design a generative probabilistic model to generate importance scores for each edge based on the input; to prune the model parameters, it views the weight's magnitude as their importance scores. Then we design an iterative co-pruning strategy to trim the graph edges and GNN weights based on their importance scores. Although it might be strikingly simple, ICPG surpasses the existing pruning method and can be universally applicable in both inductive and transductive learning settings. On ten graph-classification and two node-classification benchmarks, ICPG achieves the same performance level with $14.26\%\sim43.12\%$ sparsity for graphs and $48.80\%\sim91.41\%$ sparsity for the model.

****

Distributionally Robust Models with Parametric Likelihood Ratios

Paul Michel,Tatsunori Hashimoto,Graham Neubig

As machine learning models are deployed ever more broadly, it becomes increasingly important that they are not only able to perform well on their training distribution, but also yield accurate predictions when confronted with distribution shift. The Distributionally Robust Optimization (DRO) framework proposes to address this issue by training models to minimize their expected risk under a collection of distributions, to imitate test-time shifts. This is most commonly achieved by instance-level re-weighting of the training objective to emulate the likelihood ratio with possible test distributions, which allows for estimating their empirical risk via importance sampling (assuming that they are subpopulations of the training distribution). However, re-weighting schemes in the literature are usually limited due to the difficulty of keeping the optimization problem tractable and the complexity of enforcing normalization constraints. In this paper, we show that three simple ideas -- mini-batch level normalization, a KL penalty and simultaneous gradient updates -- allow us to train models with DRO using a broader class of parametric likelihood ratios. In a series of experiments on both image and text classification benchmarks, we find that models trained with the resulting parametric adversaries are consistently more robust to subpopulation shifts when compared to other DRO approaches, and that the method performs reliably well with little hyper-parameter tuning.

****

Neuron-Enhanced Autoencoder based Collaborative filtering: Theory and Practice
Jicong Fan,Rui Chen,Chris Ding

This paper presents a novel recommendation method called neuron-enhanced autoencoder based collaborative filtering (NE-AECF). The method uses an additional neural network to enhance the reconstruction capability of autoencoder. Different from the main neural network implemented in a layer-wise manner, the additional neural network is implemented in an element-wise manner. They are trained simultaneously to construct an enhanced autoencoder of which the activation function in the output layer is learned adaptively to approximate possibly complicated response functions in real data. We provide theoretical analysis for NE-AECF to investigate the generalization ability of autoencoder and deep learning in collaborative filtering. We prove that the element-wise neural network is able to reduce the upper bound of the prediction error for the unknown ratings, the data sparsity is not problematic but useful, and the prediction performance is closely related to the difference between the number of users and the number of items.

Numerical results show that our NE-AECF has promising performance on a few benchmark datasets.

****

Improved Image Generation via Sparsity
Roy Ganz,Michael Elad

The interest of the deep learning community in image synthesis has grown massively in recent years. Nowadays, deep generative methods, and especially Generative Adversarial Networks (GANs), are leading to state-of-the-art performance, capable of synthesizing images that appear realistic. While the efforts for improving the quality of the generated images are extensive, most attempts still consider the generator part as an uncorroborated ``black-box''. In this paper, we aim to provide a better understanding and design of the image generation process. We interpret existing generators as implicitly relying on sparsity-inspired models. More specifically, we show that generators can be viewed as manifestations of the Convolutional Sparse Coding (CSC) and its Multi-Layered version (ML-CSC) synthesis processes. We leverage this observation by explicitly enforcing a sparsifying regularization on appropriately chosen activation layers in the generator, and demonstrate that this leads to improved image synthesis. Furthermore, we show that the same rationale and benefits apply to generators serving inverse problems, demonstrated on the Deep Image Prior (DIP) method.

****

BIGRoC: Boosting Image Generation via a Robust Classifier
Roy Ganz,Michael Elad

The interest of the machine learning community in image synthesis has grown significantly in recent years, with the introduction of a wide range of deep generative models and means for training them. Such machines’ ultimate goal is to match the distributions of the given training images and the synthesized ones. In this work, we propose a general model-agnostic technique for improving the image quality and the distribution fidelity of generated images, obtained by any generative model. Our method, termed BIGRoC (boosting image generation via a robust classifier), is based on a post-processing procedure via the guidance of a given robust classifier and without a need for additional training of the generative model. Given a synthesized image, we propose to update it through projected gradient steps over the robust classifier, in an attempt to refine its recognition. We demonstrate this post-processing algorithm on various image synthesis methods and show a significant improvement of the generated images, both quantitatively and qualitatively.

****

Balancing Average and Worst-case Accuracy in Multitask Learning
Paul Michel,Sebastian Ruder,Dani Yogatama

When training and evaluating machine learning models on a large number of tasks, it is important to not only look at average task accuracy---which may be biased by easy or redundant tasks---but also worst-case accuracy (i.e. the performance on the task with the lowest accuracy). In this work, we show how to use techniq ues from the distributionally robust optimization (DRO) literature to improve worst-case performance in multitask learning. We highlight several failure cases of DRO when applied off-the-shelf and present an improved method, Lookahead-DRO (L-DRO), which mitigates these issues. The core idea of L-DRO is to anticipate the interaction between tasks during training in order to choose a dynamic re-weighting of the various task losses, which will (i) lead to minimal worst-case loss and (ii) train on as many tasks as possible. After demonstrating the efficacy of L-DRO on a small controlled synthetic setting, we evaluate it on two realistic benchmarks: a multitask version of the CIFAR-100 image classification dataset and a large-scale multilingual language modeling experiment. Our empirical results show that L-DRO achieves a better trade-off between average and worst-case accuracy with little computational overhead compared to several strong baselines.

****

Constrained Physical-Statistics Models for Dynamical System Identification and Prediction

Jérémie DONA,Marie Déchelle,patrick gallinari,Marina Levy

Modeling dynamical systems combining prior physical knowledge and machine learning (ML) is promising in scientific problems when the underlying processes are not fully understood, e.g. when the dynamics is partially known. A common practice to identify the respective parameters of the physical and ML components is to formulate the problem as supervised learning on observed trajectories. However, this formulation leads to an infinite number of possible decompositions. To solve this ill-posedness, we reformulate the learning problem by introducing an upper bound on the prediction error of a physical-statistical model. This allows us to control the contribution of both the physical and statistical components to the overall prediction. This framework generalizes several existing hybrid schemes proposed in the literature. We provide theoretical guarantees on the well-posedness of our formulation along with a proof of convergence in a simple affine setting. For more complex dynamics, we validate our framework experimentally.

****

Doubly Adaptive Scaled Algorithm for Machine Learning Using Second-Order Information

Majid Jahani,Sergey Rusakov,Zheng Shi,Peter Richtárik,Michael W. Mahoney,Martin Takac

We present a novel adaptive optimization algorithm for large-scale machine learning problems. Equipped with a low-cost estimate of local curvature and Lipschitz smoothness, our method dynamically adapts the search direction and step-size. The search direction contains gradient information preconditioned by a well-scaled diagonal preconditioning matrix that captures the local curvature information. Our methodology does not require the tedious task of learning rate tuning, as the learning rate is updated automatically without adding an extra hyper-parameter. We provide convergence guarantees on a comprehensive collection of optimization problems, including convex, strongly convex, and nonconvex problems, in both deterministic and stochastic regimes. We also conduct an extensive empirical evaluation on standard machine learning problems, justifying our algorithm's versatility and demonstrating its strong performance compared to other start-of-the-art first-order and second-order methods.

****

Towards Learning to Speak and Hear Through Multi-Agent Communication over a Continuous Acoustic Channel

Kevin Michael Eloff,Arnu Pretorius,Okko Räsänen,Herman Arnold Engelbrecht,Herman Kamper

While multi-agent reinforcement learning has been used as an effective means to study emergent communication between agents, existing work has focused almost exclusively on communication with discrete symbols. Human communication often takes place (and emerged) over a continuous acoustic channel; human infants acquire language in large part through continuous signalling with their caregivers. We therefore ask: Are we able to observe emergent language between agents with a continuous communication channel trained through reinforcement learning? And if so, what is the impact of channel characteristics on the emerging language? We prop ose an environment and training methodology to serve as a means to carry out an initial exploration of these questions. We use a simple messaging environment where a "speaker" agent needs to convey a concept to a "listener". The Speaker is equipped with a vocoder that maps symbols to a continuous waveform, this is passed over a lossy continuous channel, and the Listener needs to map the continuous signal to the concept. Using deep Q-learning, we show that basic compositionality emerges in the learned language representations. We find that noise is essential in the communication channel when conveying unseen concept combinations. And we show that we can ground the emergent communication by introducing a caregiver predisposed to "hearing" or "speaking" English. Finally, we describe how our platform serves as a starting point for future work that uses a combination of deep reinforcement learning and multi-agent systems to study our questions of continuous signalling in language learning and emergence.

****

PARS: PSEUDO-LABEL AWARE ROBUST SAMPLE SELECTION FOR LEARNING WITH NOISY LABELS

Arushi Goel,Yunlong Jiao,Jordan Massiah

Acquiring accurate labels on large-scale datasets is both time consuming and expensive. To reduce the dependency of deep learning models on learning from clean labeled data, several recent research efforts are focused on learning with noisy labels. These methods typically fall into three design categories to learn a noise robust model: sample selection approaches, noise robust loss functions, or label correction methods. In this paper, we propose PARS: Pseudo-Label Aware Robust Sample Selection, a hybrid approach that combines the best from all three worlds in a joint-training framework to achieve robustness to noisy labels. Specifically, PARS exploits all training samples using both the raw/noisy labels and estimated/refurbished pseudo-labels via self-training, divides samples into an ambiguous and a noisy subset via loss analysis, and designs label-dependent noise-aware loss functions for both sets of filtered labels. Results show that PARS significantly outperforms the state of the art on extensive studies on the noisy CIFAR-10 and CIFAR-100 datasets, particularly on challenging high-noise and low-resource settings. In particular, PARS achieved an absolute 12% improvement in test accuracy on the CIFAR-100 dataset with 90% symmetric label noise, and an absolute 27% improvement in test accuracy when only 1/5 of the noisy labels are available during training as an additional restriction. On a real-world noisy dataset, Clothing1M, PARS achieves competitive results to the state of the art.

****

Interactive Model with Structural Loss for Language-based Abductive Reasoning

Linhao Li,Ming Xu,Yongfeng Dong,Xin Li,Jianhua Tao,Qinghua Hu

The abductive natural language inference task (αNLI) is proposed to infer the most plausible explanation between the cause and the event. In the αNLI task, two observations are given, and the most plausible hypothesis is asked to pick out from the candidates. Existing methods model the relation between each candidate hypothesis separately and penalize the inference network uniformly. In this paper, we argue that it is unnecessary to distinguish the reasoning abilities among correct hypotheses; and similarly, all wrong hypotheses contribute the same when explaining the reasons of the observations. Therefore, we propose to group instead of ranking the hypotheses and design a structural loss called "joint softmax focal loss" in this paper. Based on the observation that the hypotheses are generally semantically related, we have designed a novel interactive language model aiming at exploiting the rich interaction among competing hypotheses. We name this new model for αNLI: Interactive Model with Structural Loss (IMSL). The experimental results show that our IMSL has achieved the highest performance on the RoBERTa-large pretrained model, with ACC and AUC results increased by about 1% and 5% respectively.

****

Model Compression via Symmetries of the Parameter Space

Iordan Ganev,Robin Walters

We provide a theoretical framework for neural networks in terms of the representation theory of quivers, thus revealing symmetries of the parameter space of neural networks. An exploitation of these symmetries leads to a model compression a lgorithm for radial neural networks based on an analogue of the QR decomposition. The algorithm is lossless; the compressed model has the same feedforward function as the original model. If applied before training, optimization of the compressed model by gradient descent is equivalent to a projected version of gradient descent on the original model.

****

CAGE: Probing Causal Relationships in Deep Generative Models

Joey Bose,Ricardo Pio Monti,Aditya Grover

Deep generative models excel at generating complex, high-dimensional data, often exhibiting impressive generalization beyond the training distribution. The learning principle for these models is however purely based on statistical objectives and it is unclear to what extent such models have internalized the causal relationships present in the training data, if at all. With increasing real-world deployments, such a causal understanding of generative models is essential for interpreting and controlling their use in high-stake applications that require synthetic data generation. We propose CAGE, a framework for inferring the cause-effect relationships governing deep generative models. CAGE employs careful geometrical manipulations within the latent space of a generative model for generating counterfactuals and estimating unit-level generative causal effects. CAGE does not require any modifications to the training procedure and can be used with any existing pretrained latent variable model. Moreover, the pretraining can be completely unsupervised and does not require any treatment or outcome labels. Empirically, we demonstrate the use of CAGE for: (a) inferring cause-effect relationships within a deep generative model trained on both synthetic and high resolution images, and (b) guiding data augmentations for robust classification where CAGE achieves improvements over current default approaches on image datasets.

****

Understanding approximate and unrolled dictionary learning for pattern recovery

Benoît Malézieux,Thomas Moreau,Matthieu Kowalski

Dictionary learning consists of finding a sparse representation from noisy data and is a common way to encode data-driven prior knowledge on signals. Alternating minimization (AM) is standard for the underlying optimization, where gradient descent steps alternate with sparse coding procedures. The major drawback of this method is its prohibitive computational cost, making it unpractical on large real-world data sets. This work studies an approximate formulation of dictionary learning based on unrolling and compares it to alternating minimization to find the best trade-off between speed and precision. We analyze the asymptotic behavior and convergence rate of gradients estimates in both methods. We show that unrolling performs better on the support of the inner problem solution and during the first iterations. Finally, we apply unrolling on pattern learning in magnetoencephalography (MEG) with the help of a stochastic algorithm and compare the performance to a state-of-the-art method.

****

Constraining Linear-chain CRFs to Regular Languages

Sean Papay,Roman Klinger,Sebastian Pado

A major challenge in structured prediction is to represent the interdependencies within output structures. When outputs are structured as sequences, linear-chain conditional random fields (CRFs) are a widely used model class which can learn local dependencies in the output. However, the CRF's Markov assumption makes it impossible for CRFs to represent distributions with nonlocal dependencies, and standard CRFs are unable to respect nonlocal constraints of the data (such as global arity constraints on output labels). We present a generalization of CRFs that can enforce a broad class of constraints, including nonlocal ones, by specifying the space of possible output structures as a regular language $\mathcal{L}$. The resulting regular-constrained CRF (RegCCRF) has the same formal properties as a standard CRF, but assigns zero probability to all label sequences not in $\mathcal{L}$. Notably, RegCCRFs can incorporate their constraints during training, while related models only enforce constraints during decoding. We prove that constrained training is never worse than constrained decoding, and show empirically that it can be substantially better in practice. Additionally, we demon strate a practical benefit on downstream tasks by incorporating a RegCCRF into a deep neural model for semantic role labeling, exceeding state-of-the-art results on a standard dataset.

****

Dive Deeper Into Integral Pose Regression

Kerui Gu,Linlin Yang,Angela Yao

Integral pose regression combines an implicit heatmap with end-to-end training for human body and hand pose estimation. Unlike detection-based heatmap methods, which decode final joint positions from the heatmap with a non-differentiable argmax operation, integral regression methods apply a differentiable expectation operation. This paper offers a deep dive into the inference and back-propagation of integral pose regression to better understand the differences in performance and training compared to detection-based methods. For inference, we give theoretical support as to why expectation should always be better than the argmax operation, i.e. integral regression should always outperform detection. Yet, in practice, this is observed only in hard cases because the heatmap activation for regression shrinks in easy cases. We then experimentally show that activation shrinkage is one of the leading causes for integral regression's inferior performance. For back-propagation, we theoretically and empirically analyze the gradients to explain the slow training speed of integral regression. Based on these findings, we incorporate the supervision of a spatial prior to speed up training and improve performance.

****

Invariance-Guided Feature Evolution for Few-Shot Learning

Wenming Cao,Zhineng Zhao,Qifan Liu,Zhihai He

Few-shot learning (FSL) aims to characterize the inherent visual relationship between support and query samples which can be well generalized to unseen classes so that we can accurately infer the labels of query samples from very few support samples. We observe that, in a successfully learned FSL model, this visual relationship and the learned features of the query samples should remain largely invariant across different configurations of the support set. Driven by this observation, we propose to construct a feature evolution network with an ensemble of few-shot learners evolving along different configuration dimensions. We choose to study two major parameters that control the support set configuration: the number of labeled samples per class (called shots) and the percentage of training samples (called partition) in the support set. Based on this network, we characterize and track the evolution behavior of learned query features across different shots-partition configurations, which will be minimized by a set of invariance loss functions during the training stage. Our extensive experimental results demonstrate that the proposed invariance-guided feature evolution (IGFE) method significantly improves the performance and generalization capability of few-shot learning and outperforms the state-of-the-art methods by large margins, especially in cross-domain classification tasks for generalization capability test. For example, in the cross-domain test on the fine-grained CUB image classification task, our method has improved the classification accuracy by more than 5%.

****

Local Reweighting for Adversarial Training

Ruize Gao,Feng Liu,Kaiwen Zhou,Gang Niu,Bo Han,James Cheng

Instances-reweighted adversarial training (IRAT) can significantly boost the robustness of trained models, where data being less/more vulnerable to the given attack are assigned smaller/larger weights during training. However, when tested on attacks different from the given attack simulated in training, the robustness may drop significantly (e.g., even worse than no reweighting). In this paper, we study this problem and propose our solution--locally reweighted adversarial training (LRAT). The rationale behind IRAT is that we do not need to pay much attention to an instance that is already safe under the attack. We argue that the safeness should be attack-dependent, so that for the same instance, its weight can change given different attacks based on the same model. Thus, if the attack simulated in training is mis-specified, the weights of IRAT are misleading. To this end, LRAT pairs each instance with its adversarial variants and performs local r

eweighting inside each pair, while performing no global reweighting--the rationale is to fit the instance itself if it is immune to the attack, but not to skip the pair, in order to passively defend different attacks in future. Experiments show that LRAT works better than both IRAT (i.e., global reweighting) and the standard AT (i.e., no reweighting) when trained with an attack and tested on different attacks.
****

Discovering the neural correlate informed nosological relation among multiple neuropsychiatric disorders through dual utilisation of diagnostic information
Wenjun Bai,Tomoki Tokuda,Okito Yamashita,Junichiro Yoshimoto
The unravelled nosological relation among diverse types of neuropsychiatric disorders serves as an important precursor in advocating the dimensional approach to psychiatric classification. Leveraging high-dimensional abnormal resting-state functional connectivity, the crux of mining corresponded nosological relations is to derive a low-dimensional embedding space that preserves the diagnostic attributes of represented disorders. To accomplish this goal, we seek to exploit the available diagnostic information in learning the optimal embedding space by proposing a novel type of conditional variational auto-encoder that incorporates dual utilisation of diagnostic information. Encouraged by the achieved promising results in challenging the conventional approaches in low dimensional density estimation of synthetic functional connectivity features, we further implement our approach on two empirical neuropsychiatric neuroimaging datasets and discover a reliable nosological relation among autism spectrum disorder, major depressive disorder, and schizophrenia.
****

Continuous Control with Action Quantization from Demonstrations
Robert Dadashi,Leonard Hussenot,Damien Vincent,Sertan Girgin,Anton Raichuk,Matthieu Geist,Olivier Pietquin
In Reinforcement Learning (RL), discrete actions, as opposed to continuous actions, result in less complex exploration problem and the immediate derivation of the maximum of the action-value function which is central to dynamic programming-based methods. In this paper, we propose a novel method: Action Quantization from Demonstrations (AQuaDem) to learn a discretization of continuous action spaces by leveraging the priors of demonstrations. This dramatically reduces the exploration problem, since the actions faced by the agent not only are in a finite number but also are plausible in light of the demonstrator’s behavior. By discretizing the action space we can apply any discrete action deep RL algorithm to the continuous control problem. We evaluate the proposed method on three different setups: RL with demonstrations, RL with play data --demonstrations of a human playing in an environment but not solving any specific task-- and Imitation Learning. For all three setups, we only consider human data, thus most challenging than synthetic data. We found that AQuaDem consistently outperforms state-of-the-art continuous control methods, both in terms of performance and sample efficiency.
****

Self-Contrastive Learning
Sangmin Bae,Sungnyun Kim,Jongwoo Ko,Gihun Lee,SeungJong Noh,Se-Young Yun
This paper proposes a novel contrastive learning framework, called Self-Contrastive (SelfCon) Learning, that self-contrasts within multiple outputs from the different levels of a multi-exit network. SelfCon learning does not require additional augmented samples, which resolves the concerns of multi-viewed batch (e.g., high computational cost and generalization error). Furthermore, we prove that SelfCon loss guarantees the lower bound of label-conditional mutual information between the intermediate and the last feature. In our experiments including ImageNet-100, SelfCon surpasses cross-entropy and Supervised Contrastive (SupCon) learning without the need for a multi-viewed batch. We demonstrate that the success of SelfCon learning is related to the regularization effect associated with the single-view and sub-network.
****

Personalized Neural Architecture Search for Federated Learning

Minh Hoang,Carl Kingsford

Federated Learning (FL) is a recently proposed learning paradigm for decentralized devices to collaboratively train a predictive model without exchanging private data. Existing FL frameworks, however, assume a one-size-fit-all model architecture to be collectively trained by local devices, which is determined prior to observing their data. Even with good engineering acumen, this often falls apart when local tasks are different and require diverging choices of architecture modelling to learn effectively. This motivates us to develop a novel personalized neural architecture search (NAS) algorithm for FL. Our algorithm, FedPNAS, learns a base architecture that can be structurally personalized for quick adaptation to each local task. We empirically show that FedPNAS significantly outperforms other NAS and FL benchmarks on several real-world datasets.

****

Near-Optimal Reward-Free Exploration for Linear Mixture MDPs with Plug-in Solver

Xiaoyu Chen,Jiachen Hu,Lin Yang,Liwei Wang

Although model-based reinforcement learning (RL) approaches are considered more sample efficient, existing algorithms are usually relying on sophisticated planning algorithm to couple tightly with the model-learning procedure. Hence the learned models may lack the ability of being re-used with more specialized planners. In this paper we address this issue and provide approaches to learn an RL model efficiently without the guidance of a reward signal. In particular, we take a plug-in solver approach, where we focus on learning a model in the exploration phase and demand that \emph{any planning algorithm} on the learned model can give a near-optimal policy. Specicially, we focus on the linear mixture MDP setting, where the probability transition matrix is a (unknown) convex combination of a set of existing models. We show that, by establishing a novel exploration algorithm, the plug-in approach learns a model by taking $\tilde{O}(d^2H^3/\epsilon^2)$ interactions with the environment and \emph{any} ϵ-optimal planner on the model gives an $O(\epsilon)$-optimal policy on the original model. This sample complexity matches lower bounds for non-plug-in approaches and is \emph{statistically optimal}. We achieve this result by leveraging a careful maximum total-variance bound using Bernstein inequality and properties specified to linear mixture MDP.

****

Evidential Turing Processes

Melih Kandemir,Abdullah Akgül,Manuel Haussmann,Gozde Unal

A probabilistic classifier with reliable predictive uncertainties i) fits successfully to the target domain data, ii) provides calibrated class probabilities in difficult regions of the target domain (e.g. class overlap), and iii) accurately identifies queries coming out of the target domain and reject them. We introduce an original combination of Evidential Deep Learning, Neural Processes, and Neural Turing Machines capable of providing all three essential properties mentioned above for total uncertainty quantification. We observe our method on three image classification benchmarks to consistently improve the in-domain uncertainty quantification, out-of-domain detection, and robustness against input perturbations with one single model. Our unified solution delivers an implementation-friendly and computationally efficient recipe for safety clearance and provides intellectual economy to an investigation of algorithmic roots of epistemic awareness in deep neural nets.

****

Noisy Feature Mixup

Soon Hoe Lim,N. Benjamin Erichson,Francisco Utrera,Winnie Xu,Michael W. Mahoney

We introduce Noisy Feature Mixup (NFM), an inexpensive yet effective method for data augmentation that combines the best of interpolation based training and noise injection schemes. Rather than training with convex combinations of pairs of examples and their labels, we use noise-perturbed convex combinations of pairs of data points in both input and feature space. This method includes mixup and manifold mixup as special cases, but it has additional advantages, including better smoothing of decision boundaries and enabling improved model robustness. We provide theory to understand this as well as the implicit regularization effects o f NFM. Our theory is supported by empirical results, demonstrating the advantage of NFM, as compared to mixup and manifold mixup. We show that residual networks and vision transformers trained with NFM have favorable trade-offs between predictive accuracy on clean data and robustness with respect to various types of data perturbation across a range of computer vision benchmark datasets.

****

Structure by Architecture: Disentangled Representations without Regularization

Felix Leeb,Giulia Lanzillotta,Yashas Annadani,Michel Besserve,Stefan Bauer,Bernhard Schölkopf

We study the problem of self-supervised structured representation learning using autoencoders for downstream tasks such as generative modeling. Unlike most methods which rely on matching an arbitrary, relatively unstructured, prior distribution for sampling, we propose a sampling technique that relies solely on the independence of latent variables, thereby avoiding the trade-off between reconstruction quality and generative performance inherent to VAEs. We design a novel autoencoder architecture capable of learning a structured representation without the need for aggressive regularization. Our structural decoders learn a hierarchy of latent variables, akin to structural causal models, thereby ordering the information without any additional regularization. We demonstrate how these models learn a representation that improves results in a variety of downstream tasks including generation, disentanglement, and extrapolation using several challenging and natural image datasets.

****

Deep learning via message passing algorithms based on belief propagation

Fabrizio Pittorino,Carlo Lucibello,Gabriele Perugini,Riccardo Zecchina

Message-passing algorithms based on the Belief Propagation (BP) equations constitute a well-known distributed computational scheme. It is exact on tree-like graphical models and has also proven to be effective in many problems defined on graphs with loops (from inference to optimization, from signal processing to clustering).

The BP-based scheme is fundamentally different from stochastic gradient descent (SGD), on which the current success of deep networks is based. In this paper, we present and adapt to mini-batch training on GPUs a family of BP-based message-passing algorithms with a reinforcement field that biases distributions towards locally entropic solutions.

These algorithms are capable of training multi-layer neural networks with discrete weights and activations with performance comparable to SGD-inspired heuristics (BinaryNet) and are naturally well-adapted to continual learning. Furthermore, using these algorithms to estimate the marginals of the weights allows us to make approximate Bayesian predictions that have higher accuracy than point-wise solutions.

****

Adam is no better than normalized SGD: Dissecting how adaptivity improves GAN performance

Samy Jelassi,Arthur Mensch,Gauthier Gidel,Yuanzhi Li

Adaptive methods are widely used for training generative adversarial networks (GAN). While there has been some work to pinpoint the marginal value of adaptive methods in minimization problems, it remains unclear why it is still the method of choice for GAN training. This paper formally studies how adaptive methods help performance in GANs. First, we dissect Adam---the most popular adaptive method for GAN training---by comparing with SGDA the direction and the norm of its update vector. We empirically show that SGDA with the same vector norm as Adam reaches similar or even better performance than the latter. This empirical study encourages us to consider normalized stochastic gradient descent ascent (nSGDA) as a simpler alternative to Adam. We then propose a synthetic theoretical framework to understand why nSGDA yields better performance than SGDA for GANs. In that situation, we prove that a GAN trained with nSGDA provably recovers all the modes of the true distribution. In contrast, the same networks trained with SGDA (and any learning rate configuration) suffers from mode collapsing. The critical insight in our analysis is that normalizing the gradients forces the discriminato r and generator to update at the same pace. We empirically show the competitive performance of nSGDA on real-world datasets.

****

Peek-a-Boo: What (More) is Disguised in a Randomly Weighted Neural Network, and How to Find It Efficiently

Xiaohan Chen,Jason Zhang,Zhangyang Wang

Sparse neural networks (NNs) are intensively investigated in literature due to their appeal in saving storage, memory, and computational costs. A recent work (Ramanujan et al., 2020) showed that, different from conventional pruning-and-finetuning pipeline, there exist hidden subnetworks in randomly initialized NNs that have good performance without training the weights. However, such "hidden subnetworks" have mediocre performances and require an expensive edge-popup algorithm to search for them. In this work, we define an extended class of subnetworks in randomly initialized NNs called disguised subnetworks, which are not only "hidden" in the random networks but also "disguised" -- hence can only be "unmasked" with certain transformations on weights. We argue that the unmasking process plays an important role in enlarging the capacity of the subnetworks and thus grants two major benefits: (i) the disguised subnetworks easily outperform the hidden counterparts; (ii) the unmasking process helps to relax the quality requirement on the sparse subnetwork mask so that the expensive edge-popup algorithm can be replaced with more efficient alternatives. On top of this new concept, we propose a novel two-stage algorithm that plays a Peek-a-Boo (PaB) game to identify the disguised subnetworks with a combination of two operations: (1) searching efficiently for a subnetwork at random initialization; (2) unmasking the disguise by learning to transform the resulting subnetwork's remaining weights. Furthermore, we show that the unmasking process can be efficiently implemented (a) without referring to any latent weights or scores; and (b) by only leveraging approximated gradients, so that the whole training algorithm is computationally light. Extensive experiments with several large models (ResNet-18, ResNet-50, and WideResNet-28) and datasets (CIFAR-10, CIFAR-100 and ImageNet) demonstrate the competency of PaB over edge-popup and other counterparts. Our codes are available at: https://github.com/VITA-Group/Peek-a-Boo.

****

Meta Discovery: Learning to Discover Novel Classes given Very Limited Data

Haoang Chi,Feng Liu,Wenjing Yang,Long Lan,Tongliang Liu,Bo Han,Gang Niu,Mingyuan Zhou,Masashi Sugiyama

In novel class discovery (NCD), we are given labeled data from seen classes and unlabeled data from unseen classes, and we train clustering models for the unseen classes. However, the implicit assumptions behind NCD are still unclear. In this paper, we demystify assumptions behind NCD and find that high-level semantic features should be shared among the seen and unseen classes. Based on this finding, NCD is theoretically solvable under certain assumptions and can be naturally linked to meta-learning that has exactly the same assumption as NCD. Thus, we can empirically solve the NCD problem by meta-learning algorithms after slight modifications. This meta-learning-based methodology significantly reduces the amount of unlabeled data needed for training and makes it more practical, as demonstrated in experiments. The use of very limited data is also justified by the application scenario of NCD: since it is unnatural to label only seen-class data, NCD is sampling instead of labeling in causality. Therefore, unseen-class data should be collected on the way of collecting seen-class data, which is why they are novel and first need to be clustered.

****

Defect Transfer GAN: Diverse Defect Synthesis for Data Augmentation

Ruyu Wang,Sabrina Hoppe,Eduardo Monari,Marco Huber

Large amounts of data are a common requirement for many deep learning approaches. However, data is not always equally available at large scale for all classes. For example, on highly optimized production lines, defective samples are hardly acquired while non-defective samples come almost for free. The defects however often seem to resemble each other, e.g., scratches on different products may only differ in few characteristics. In this work, we propose to make use of the shared characteristics by transferring a stylized defect-specific content from one type of background product to another. Moreover, the stochastic variations of the shared characteristics are captured, which also allows generating novel defects from random noise. These synthetic defective samples enlarge the dataset and increase the diversity of defects on the target product. Experiments demonstrate that our model is able to disentangle the defect-specific content from the background of an image without pixel-level labels. We present convincing results on images from real industrial production lines. Furthermore, we show consistent gains of using our method to enlarge training sets in classification tasks.

****

How Well Does Self-Supervised Pre-Training Perform with Streaming Data?

Dapeng Hu,Shipeng Yan,Qizhengqiu Lu,Lanqing HONG,Hailin Hu,Yifan Zhang,Zhenguo Li,Xinchao Wang,Jiashi Feng

Prior works on self-supervised pre-training focus on the joint training scenario, where massive unlabeled data are assumed to be given as input all at once, and only then is a learner trained. Unfortunately, such a problem setting is often impractical if not infeasible since many real-world tasks rely on sequential learning, e.g., data are decentralized or collected in a streaming fashion. In this paper, we conduct the first thorough and dedicated investigation on self-supervised pre-training with streaming data, aiming to shed light on the model behavior under this overlooked setup. Specifically, we pre-train over 500 models on four categories of pre-training streaming data from ImageNet and DomainNet and evaluate them on three types of downstream tasks and 12 different downstream datasets. Our studies show that, somehow beyond our expectation, with simple data replay or parameter regularization, sequential self-supervised pre-training turns out to be an efficient alternative for joint pre-training, as the performances of the former are mostly on par with those of the latter. Moreover, catastrophic forgetting, a common issue in sequential supervised learning, is much alleviated in sequential self-supervised learning (SSL), which is well justified through our comprehensive empirical analysis on representations and the sharpness of minima in the loss landscape. Our findings, therefore, suggest that, in practice, for SSL, the cumbersome joint training can be replaced mainly by sequential learning, which in turn enables a much broader spectrum of potential application scenarios.

****

Constrained Policy Optimization via Bayesian World Models

Yarden As,Ilnura Usmanova,Sebastian Curi,Andreas Krause

Improving sample-efficiency and safety are crucial challenges when deploying reinforcement learning in high-stakes real world applications. We propose LAMBDA, a novel model-based approach for policy optimization in safety critical tasks modeled via constrained Markov decision processes. Our approach utilizes Bayesian world models, and harnesses the resulting uncertainty to maximize optimistic upper bounds on the task objective, as well as pessimistic upper bounds on the safety constraints. We demonstrate LAMBDA's state of the art performance on the Safety-Gym benchmark suite in terms of sample efficiency and constraint violation.

****

Intrusion-Free Graph Mixup

Hongyu Guo,Yongyi Mao

We present a simple and yet effective interpolation-based regularization technique to improve the generalization of Graph Neural Networks (GNNs). We leverage the recent advances in Mixup regularizer for vision and text, where random sample pairs and their labels are interpolated to create synthetic samples for training. Unlike images or natural sentences, which embrace a grid or linear sequence format, graphs have arbitrary structure and topology, which play a vital role on the semantic information of a graph. Consequently, even simply deleting or adding one edge from a graph can dramatically change its semantic meanings. This makes interpolating graph inputs very challenging because mixing random graph pairs may naturally create graphs with identical structure but with different labels, causing the manifold intrusion issue. To cope with this obstacle, we propose t he first input mixing schema for Mixup on graph. We theoretically prove that our mixing strategy can recover the source graphs from the mixed graph, and guarantees that the mixed graphs are manifold intrusion free. We also empirically show that our method can effectively regularize the graph classification learning, resulting in superior predictive accuracy over popular graph augmentation baselines.

****

Reachability Traces for Curriculum Design in Reinforcement Learning

Thommen Karimpanal George,Majid Abdolshah,Hung Le,Santu Rana,Sunil Gupta,Truyen Tran,Svetha Venkatesh

The objective in goal-based reinforcement learning is to learn a policy to reach a particular goal state within the environment. However, the underlying reward function may be too sparse for the agent to efficiently learn useful behaviors. Recent studies have demonstrated that reward sparsity can be overcome by instead learning a curriculum of simpler subtasks. In this work, we design an agent's curriculum by focusing on the aspect of goal reachability, and introduce the idea of a reachability trace, which is used as a basis to determine a sequence of intermediate subgoals to guide the agent towards its primary goal. We discuss several properties of the trace function, and in addition, validate our proposed approach empirically in a range of environments, while comparing its performance against appropriate baselines.

****

Subspace Regularizers for Few-Shot Class Incremental Learning

Afra Feyza Akyürek,Ekin Akyürek,Derry Wijaya,Jacob Andreas

Few-shot class incremental learning---the problem of updating a trained classifier to discriminate among an expanded set of classes with limited labeled data---is a key challenge for machine learning systems deployed in non-stationary environments. Existing approaches to the problem rely on complex model architectures and training procedures that are difficult to tune and re-use. In this paper, we present an extremely simple approach that enables the use of ordinary logistic regression classifiers for few-shot incremental learning. The key to this approach is a new family of \textit{subspace regularization} schemes that encourage weight vectors for new classes to lie close to the subspace spanned by the weights of existing classes. When combined with pretrained convolutional feature extractors, logistic regression models trained with subspace regularization outperform specialized, state-of-the-art approaches to few-shot incremental image classification by up to 23\% on the \textit{mini}ImageNet dataset. Because of its simplicity, subspace regularization can be straightforwardly configured to incorporate additional background information about the new classes (including class names and descriptions specified in natural language); this offers additional control over the trade-off between existing and new classes. Our results show that simple geometric regularization of class representations offers an effective tool for continual learning.

****

Using Graph Representation Learning with Schema Encoders to Measure the Severity of Depressive Symptoms

Simin Hong,Anthony Cohn,David Crossland Hogg

Graph neural networks (GNNs) are widely used in regression and classification problems applied to text, in areas such as sentiment analysis and medical decision-making processes. We propose a novel form for node attributes within a GNN based model that captures node-specific embeddings for every word in the vocabulary. This provides a global representation at each node, coupled with node-level updates according to associations among words in a transcript. We demonstrate the efficacy of the approach by augmenting the accuracy of measuring major depressive disorder (MDD). Prior research has sought to make a diagnostic prediction of depression levels from patient data using several modalities, including audio, video, and text. On the DAIC-WOZ benchmark, our method outperforms state-of-art methods by a substantial margin, including those using multiple modalities. Moreover, we also evaluate the performance of our novel model on a Twitter sentiment dataset. We show that our model outperforms a general GNN model by leveraging our

novel 2-D node attributes. These results demonstrate the generality of the proposed method.
****
Actor-Critic Policy Optimization in a Large-Scale Imperfect-Information Game
Haobo Fu,Weiming Liu,Shuang Wu,Yijia Wang,Tao Yang,Kai Li,Junliang Xing,Bin Li,Bo Ma,QIANG FU,Yang Wei
The deep policy gradient method has demonstrated promising results in many large-scale games, where the agent learns purely from its own experience. Yet, policy gradient methods with self-play suffer convergence problems to a Nash Equilibrium (NE) in multi-agent situations. Counterfactual regret minimization (CFR) has a convergence guarantee to a NE in 2-player zero-sum games, but it usually needs domain-specific abstractions to deal with large-scale games. Inheriting merits from both methods, in this paper we extend the actor-critic algorithm framework in deep reinforcement learning to tackle a large-scale 2-player zero-sum imperfect-information game, 1-on-1 Mahjong, whose information set size and game length are much larger than poker. The proposed algorithm, named Actor-Critic Hedge (ACH), modifies the policy optimization objective from originally maximizing the discounted returns to minimizing a type of weighted cumulative counterfactual regret. This modification is achieved by approximating the regret via a deep neural network and minimizing the regret via generating self-play policies using Hedge. ACH is theoretically justified as it is derived from a neural-based weighted CFR, for which we prove the convergence to a NE under certain conditions. Experimental results on the proposed 1-on-1 Mahjong benchmark and benchmarks from the literature demonstrate that ACH outperforms related state-of-the-art methods. Also, the agent obtained by ACH defeats a human champion in 1-on-1 Mahjong.
****
Towards understanding how momentum improves generalization in deep learning
Samy Jelassi,Yuanzhi Li
Stochastic gradient descent (SGD) with momentum is widely used for training modern deep learning architectures. While it is well understood that using momentum can lead to faster convergence rate in various settings, it has also been observed that momentum yields higher generalization. Prior work argue that momentum stabilizes the SGD noise during training and this leads to higher generalization. In this paper, we take the opposite view to this result and first empirically show that gradient descent with momentum (GD+M) significantly improves generalization comparing to gradient descent (GD) in many deep learning tasks. From this observation, we formally study how momentum improves generalization in deep learning. We devise a binary classification setting where a two-layer (over-parameterized) convolutional neural network trained with GD+M provably generalizes better than the same network trained with vanilla GD, when both algorithms start from the same random initialization. The key insight in our analysis is that momentum is beneficial in datasets where the examples share some features but differ in their margin. Contrary to the GD model that memorizes the small margin data, GD+M can still learn the features in these data thanks to its historical gradients. We also empirically verify this learning process of momentum in real-world settings.
****
Policy Gradients Incorporating the Future
David Venuto,Elaine Lau,Doina Precup,Ofir Nachum
Reasoning about the future -- understanding how decisions in the present time affect outcomes in the future -- is one of the central challenges for reinforcement learning (RL), especially in highly-stochastic or partially observable environments. While predicting the future directly is hard, in this work we introduce a method that allows an agent to ``look into the future'' without explicitly predicting it. Namely, we propose to allow an agent, during its training on past experience, to observe what \emph{actually} happened in the future at that time, while enforcing an information bottleneck to avoid the agent overly relying on this privileged information. Coupled with recent advances in variational inference and a latent-variable autoregressive model, this gives our agent the ability to utilize rich and \emph{useful} information about the future trajectory dynamics

in addition to the present. Our method, Policy Gradients Incorporating the Future (PGIF), is easy to implement and versatile, being applicable to virtually any policy gradient algorithm. We apply our proposed method to a number of off-the-shelf RL algorithms and show that PGIF is able to achieve higher reward faster in a variety of online and offline RL domains, as well as sparse-reward and partially observable environments.

****

GSmooth: Certified Robustness against Semantic Transformations via Generalized Randomized Smoothing

Hao Zhongkai,Chengyang Ying,Yinpeng Dong,Hang Su,Jun Zhu

The vulnerability of deep learning models to adversarial examples and semantic transformations has limited the applications in risk-sensitive areas. The recent development of certified defense approaches like randomized smoothing provides a promising direction towards building reliable machine learning systems. However, current certified defenses cannot handle complex semantic transformations like rotational blur and defocus blur which are common in practical applications. In this paper, we propose a generalized randomized smoothing framework (GSmooth) for certified robustness against semantic transformations. We provide both a unified and rigorous theoretical framework and scalable algorithms for certified robustness on complex semantic transformations. Specifically, our key idea is to use a surrogate image-to-image neural network to approximate a transformation which provides a powerful tool for studying the properties of semantic transformations and certify the transformation based on this neural network. Experiments on multiple types of semantic perturbations and corruptions using multiple datasets demonstrate the effectiveness of our approach.

****

Disentangled generative models for robust dynamical system prediction

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

Deep neural networks have become increasingly of interest in dynamical system prediction, but out-of-distribution generalization and long-term stability still remains challenging. In this work, we treat the domain parameters of dynamical systems as factors of variation of the data generating process. By leveraging ideas from supervised disentanglement and causal factorization, we aim to separate the domain parameters from the dynamics in the latent space of generative models. In our experiments we model dynamics both in phase space and in video sequences and conduct rigorous OOD evaluations. Results indicate that disentangled models adapt better to domain parameters spaces that were not present in the training data while, at the same time, provide better long-term predictions in video sequences.

****

Truth Table Deep Convolutional Neural Network, A New SAT-Encodable Architecture - Application To Complete Robustness

Adrien Benamira,Thomas Peyrin,Bryan Hooi

With the expanding role of neural networks, the need for formal verification of their behavior, interpretability and human post-processing has become critical in many applications. In 2018, it has been shown that Binary Neural Networks (BNNs) have an equivalent representation in boolean logic and can be formally analyzed using logical reasoning tools such as SAT or MaxSAT solvers. This formulation is powerful as it allows us to address a vast range of questions: existential, probabilistic, explanation generation, etc. However, to date, only BNNs can be transformed into a SAT formula and their strong binary constraints limit their natural accuracy. Moreover, the corresponding SAT conversion method intrinsically leads to formulas with a large number of variables and clauses, impeding interpretability as well as formal verification scalability. In this work, we introduce Truth Table Deep Convolutional Neural Networks (TT-DCNNs), a new family of SAT-encodable models featuring real-valued weights and real intermediate values as well as a highly interpretable conversion method. The TT-DCNN architecture enables for the first time all the logical classification rules to be extracted from a

performant neural network which can be then easily interpreted by anyone familiar with the domain. Therefore, this allows integrating human knowledge in post-processing as well as enumerating all possible inputs/outputs prior to deployment in production. We believe our new architecture paves the way between eXplainability AI (XAI) and formal verification. First, we experimentally show that TT-DCNNs offer a better tradeoff between natural accuracy and formal verification than BNNs. Then, in the robustness verification setting, we demonstrate that TT-DCNNs outperform the verifiable accuracy of BNNs with a comparable computation time. Finally, we also drastically decrease the number of clauses and variables, enabling the usage of general SAT solvers and exact model counting solvers. Our developed real-valued network has general applications and we believe that its demonstrated robustness constitutes a suitable response to the rising demand for functional formal verification.
****

Online approximate factorization of a kernel matrix by a Hebbian neural network
Kyle Luther,Sebastian Seung

We derive an online algorithm for unsupervised learning based on representing every input $\mathbf{x}_t$ by a high dimensional vector $\mathbf{y}_t$ with pairwise inner products that approximately match input similarities as measured by a kernel function: $\mathbf{y}_s \cdot \mathbf{y}_{t} \approx f(\mathbf{x}_s, \mathbf{x}_{t})$. The approximation is formulated using the objective function for classical multidimensional scaling. We derive an upper bound for this objective which only involves correlations between output vectors and nonlinear functions of input vectors. Minimizing this upper bound leads to a minimax optimization, which can be solved via stochastic gradient descent-ascent. This online algorithm can be interpreted as a recurrent neural network with Hebbian and anti-Hebbian connections, generalizing previous work on linear similarity matching. Through numerical experiments with two datasets, we demonstrate that unsupervised learning can be aided by the nonlinearity inherent in our kernel method. We also show that heavy-tailed representation vectors emerge from the learning even though no sparseness prior is used, lending further biological plausibility to the model. Our upper bound employs a rank-one Nystrom approximation to the kernel function, with the novelty of leading to an online algorithm that optimizes landmark placement.
****

VAE Approximation Error: ELBO and Exponential Families
Alexander Shekhovtsov,Dmitrij Schlesinger,Boris Flach

The importance of Variational Autoencoders reaches far beyond standalone generative models -- the approach is also used for learning latent representations and can be generalized to semi-supervised learning. This requires a thorough analysis of their commonly known shortcomings: posterior collapse and approximation errors. This paper analyzes VAE approximation errors caused by the combination of the ELBO objective and encoder models from conditional exponential families, including, but not limited to, commonly used conditionally independent discrete and continuous models.

We characterize subclasses of generative models consistent with these encoder families. We show that the ELBO optimizer is pulled away from the likelihood optimizer towards the consistent subset and study this effect experimentally. Importantly, this subset can not be enlarged, and the respective error cannot be decreased, by considering deeper encoder/decoder networks.
****

Gradient Information Matters in Policy Optimization by Back-propagating through Model
Chongchong Li,Yue Wang,Wei Chen,Yuting Liu,Zhi-Ming Ma,Tie-Yan Liu

Model-based reinforcement learning provides an efficient mechanism to find the optimal policy by interacting with the learned environment. In addition to treating the learned environment like a black-box simulator, a more effective way to use the model is to exploit its differentiability. Such methods require the gradient information of the learned environment model when calculating the policy gradient. However, since the error of gradient is not considered in the model learn ing phase, there is no guarantee for the model's accuracy. To address this problem, we first analyze the convergence rate for the policy optimization methods when the policy gradient is calculated using the learned environment model. The theoretical results show that the model gradient error matters in the policy optimization phrase. Then we propose a two-model-based learning method to control the prediction error and the gradient error. We separate the different roles of these two models at the model learning phase and coordinate them at the policy optimization phase. After proposing the method, we introduce the directional derivative projection policy optimization (DDPPO) algorithm as a practical implementation to find the optimal policy. Finally, we empirically demonstrate the proposed algorithm has better sample efficiency when achieving a comparable or better performance on benchmark continuous control tasks.

****

Imperceptible Black-box Attack via Refining in Salient Region

Zeyu Dai,Shengcai Liu,Ke Tang,Qing Li

Deep neural networks are vulnerable to adversarial examples, even in the black-box setting where the attacker only has query access to the model output. Recent studies have devised successful black-box attacks with high query efficiency. However, such performance often comes at the cost of the imperceptibility of adversarial attacks, which is essential for attackers. To address this issue, in this paper we propose to use segmentation priors for black-box attacks such that the perturbations are limited in the salient region. We find that state-of-the-art black-box attacks equipped with segmentation priors can achieve much better imperceptibility performance with little reduction in query efficiency and success rate. We further propose the Saliency Attack, a new gradient-free black-box attack that can further improve the imperceptibility by refining perturbations in the salient region. Experimental results show that the perturbations generated by our approach are much more imperceptible than the ones generated by other attacks, and are interpretable to some extent. Furthermore, our approach is found to be more robust to detection-based defense, which demonstrates its efficacy as well.

****

EinSteinVI: General and Integrated Stein Variational Inference

Ola Rønning,Ahmad Salim Al-Sibahi,Christophe Ley,Thomas Hamelryck

Stein variational inference is a technique for approximate Bayesian inference that has recently gained popularity because it combines the scalability of variational inference (VI) with the flexibility of non-parametric inference methods. While there has been considerable progress in developing algorithms for Stein variational inference, integration in existing probabilistic programming languages (PPLs) with an easy-to-use interface is currently lacking. EinSteinVI is a lightweight compostable library that integrates the latest Stein variational inference method with the PPL NumPyro (Phan et al., 2019). EinSteinVI provides ELBO-within-Stein to support the use of custom inference programs (guides), implementations of a wide range of kernels, non-linear scaling of the repulsion force (Wang & Liu,2019b), and second-order gradient updates using matrix-valued kernels (Wang et al.,2019b). We illustrate EinSteinVI using toy examples and show results on par with or better than existing state-of-the-art methods for real-world problems. These include Bayesian neural networks for regression and a Stein-mixture deep Markov model, which shows EinSteinVI scales to large models with more than 500,000 parameters.

****

VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning

Adrien Bardes,Jean Ponce,Yann LeCun

Recent self-supervised methods for image representation learning maximize the agreement between embedding vectors produced by encoders fed with different views of the same image. The main challenge is to prevent a collapse in which the encoders produce constant or non-informative vectors. We introduce VICReg (Variance-Invariance-Covariance Regularization), a method that explicitly avoids the collapse problem with two regularizations terms applied to both embeddings separatel y: (1) a term that maintains the variance of each embedding dimension above a threshold, (2) a term that decorrelates each pair of variables. Unlike most other approaches to the same problem, VICReg does not require techniques such as: weight sharing between the branches, batch normalization, feature-wise normalization, output quantization, stop gradient, memory banks, etc., and achieves results on par with the state of the art on several downstream tasks. In addition, we show that our variance regularization term stabilizes the training of other methods and leads to performance improvements.

****

YOUR AUTOREGRESSIVE GENERATIVE MODEL CAN BE BETTER IF YOU TREAT IT AS AN ENERGY-BASED ONE

Yezhen Wang,Tong Che,Bo Li,Kaitao Song,Hengzhi Pei,Yoshua Bengio,Dongsheng Li

Autoregressive generative models are commonly used, especially for those tasks involving sequential data. They have, however, been plagued by a slew of inherent flaws due to the intrinsic characteristics of chain-style conditional modeling (e.g., exposure bias or lack of long-range coherence), severely limiting their ability to model distributions properly. In this paper, we propose a unique method for training the autoregressive generative model that takes advantage of a well-designed energy-based learning objective. We show that our method is capable of alleviating the exposure bias problem and increase temporal coherence by imposing a constraint which fits joint distributions at each time step. Besides, unlike former energy-based models, we estimate energy scores based on the underlying autoregressive network itself, which does not require any extra network. Finally, thanks to importance sampling, we can train the entire model efficiently without requiring an MCMC process. Extensive empirical results, covering benchmarks like language modeling, neural machine translation, and image generation, demonstrate the effectiveness of the proposed approach.

****

Can Reinforcement Learning Efficiently Find Stackelberg-Nash Equilibria in General-Sum Markov Games?

Han Zhong,Zhuoran Yang,Zhaoran Wang,Michael Jordan

We study multi-player general-sum Markov games with one of the players designated as the leader and the rest regarded as the followers. In particular, we focus on the class of games where the state transitions are only determined by the leader's action while the actions of all the players determine their immediate rewards. For such a game, our goal is to find the Stackelberg-Nash equilibrium (SNE), which is a policy pair (π^*, ν^*) such that (i) π^* is the optimal policy for the leader when the followers always play their best response, and (ii) ν^* is the best response policy of the followers, which is a Nash equilibrium of the followers' game induced by π^*. We develop sample efficient reinforcement learning (RL) algorithms for solving SNE for both the online and offline settings. Respectively, our algorithms are optimistic and pessimistic variants of least-squares value iteration and are readily able to incorporate function approximation for handling large state spaces. Furthermore, for the case with linear function approximation, we prove that our algorithms achieve sublinear regret and suboptimality under online and offline setups respectively. To our best knowledge, we establish the first provably efficient RL algorithms for solving SNE in general-sum Markov games with leader-controlled state transitions.

****

High Probability Generalization Bounds with Fast Rates for Minimax Problems

Shaojie Li,Yong Liu

Minimax problems are receiving an increasing amount of attention in a wide range of applications in machine learning (ML), for instance, reinforcement learning, robust optimization, adversarial learning, and distributed computing, to mention but a few. Current studies focus on the fundamental understanding of general minimax problems with an emphasis on convergence behavior. As a comparison, there is far less work to study the generalization performance. Additionally, existing generalization bounds are almost all derived in expectation, and the high probability bounds are all presented in the slow order $\mathcal{O}(1/\sqrt{n})$, where n is the sample size. In this paper, we provide improved generalization an alyses and obtain sharper high probability generalization bounds for most existing generalization measures of minimax problems. We then use the improved learning bounds to establish high probability generalization bounds with fast rates for classical empirical saddle point (ESP) solution and several popular gradient-based optimization algorithms, including gradient descent ascent (GDA), stochastic gradient descent ascent (SGDA), proximal point method (PPM), extra-gradient (EG), and optimistic gradient descent ascent (OGDA). In summary, we provide a systematical analysis of sharper generalization bounds of minimax problems.

****

SUMNAS: Supernet with Unbiased Meta-Features for Neural Architecture Search

Hyeonmin Ha,Ji-Hoon Kim,Semin Park,Byung-Gon Chun

One-shot Neural Architecture Search (NAS) usually constructs an over-parameterized network, which we call a supernet, and typically adopts sharing parameters among the sub-models to improve computational efficiency. One-shot NAS often repeatedly samples sub-models from the supernet and trains them to optimize the shared parameters. However, this training strategy suffers from multi-model forgetting. Training a sampled sub-model overrides the previous knowledge learned by the other sub-models, resulting in an unfair performance evaluation between the sub-models. We propose Supernet with Unbiased Meta-Features for Neural Architecture Search (SUMNAS), a supernet learning strategy based on meta-learning to tackle the knowledge forgetting issue. During the training phase, we explicitly address the multi-model forgetting problem and help the supernet learn unbiased meta-features, independent from the sampled sub-models. Once training is over, sub-models can be instantly compared to get the overall ranking or the best sub-model. Our evaluation on the NAS-Bench-201 and MobileNet-based search space demonstrate that SUMNAS shows improved ranking ability and finds architectures whose performance is on par with existing state-of-the-art NAS algorithms.

****

Memory-Constrained Policy Optimization

Hung Le,Thommen Karimpanal George,Majid Abdolshah,Dung Nguyen,Kien Do,Sunil Gupta,Svetha Venkatesh

We introduce a new 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 to form a second trust region through the construction of 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 in case the old policy performs badly. More importantly, 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 as Memory-Constrained Policy Optimization (MCPO), is examined on a diverse suite of environments including robotic locomotion control, navigation with sparse rewards and Atari games, consistently demonstrating competitive performance against recent on-policy constrained policy gradient methods.

****

AI-SARAH: Adaptive and Implicit Stochastic Recursive Gradient Methods

Zheng Shi,Nicolas Loizou,Peter Richtárik,Martin Takac

We present AI-SARAH, a practical variant of SARAH. As a variant of SARAH, this algorithm employs the stochastic recursive gradient yet adjusts step-size based on local geometry. AI-SARAH implicitly computes step-size and efficiently estimates local Lipschitz smoothness of stochastic functions. It is fully adaptive, tune-free, straightforward to implement, and computationally efficient. We provide technical insight and intuitive illustrations on its design and convergence. We conduct extensive empirical analysis and demonstrate its strong performance compared with its classical counterparts and other state-of-the-art first-order methods in solving convex machine learning problems.

****

Intervention Adversarial Auto-Encoder

Yang Hu,Cheng Zhang

In this paper we propose a new method to stabilize the training process of the latent variables of adversarial auto-encoders, which we name Intervention Adversarial auto-encoder (IVAAE). The main idea is to introduce a sequence of distributions that bridge the distribution of the learned latent variable and its prior distribution. We theoretically and heuristically demonstrate that such bridge-like distributions, realized by a multi-output discriminator, have an effect on guiding the initial latent distribution towards the target one and hence stabilizing the training process. Several different types of the bridge distributions are proposed. We also apply a novel use of Stein variational gradient descent (SVGD), by which point assemble develops in a smooth and gradual fashion. We conduct experiments on multiple real-world datasets. It shows that IVAAE enjoys a more stable training process and achieves a better generating performance compared to the vanilla Adversarial auto-encoder (AAE)

****

Causal discovery from conditionally stationary time-series
Carles Balsells Rodas,Ruibo Tu,Hedvig Kjellstrom

Causal discovery, i.e., inferring underlying cause-effect relationships from observations of a scene or system, is an inherent mechanism in human cognition, but has been shown to be highly challenging to automate. The majority of approaches in the literature aiming for this task consider constrained scenarios with fully observed variables or data from stationary time-series.

In this work we aim for causal discovery in a more general class of scenarios, scenes with non-stationary behavior over time. For our purposes we here regard a scene as a composition objects interacting with each other over time. Non-stationarity is modeled as stationarity conditioned on an underlying variable, a state, which can be of varying dimension, more or less hidden given observations of the scene, and also depend more or less directly on these observations.

We propose a probabilistic deep learning approach called State-Dependent Causal Inference (SDCI) for causal discovery in such conditionally stationary time-series data. Results in two different synthetic scenarios show that this method is able to recover the underlying causal dependencies with high accuracy even in cases with hidden states.

****

Temporal Efficient Training of Spiking Neural Network via Gradient Re-weighting
Shikuang Deng,Yuhang Li,Shanghang Zhang,Shi Gu

Recently, brain-inspired spiking neuron networks (SNNs) have attracted widespread research interest because of their event-driven and energy-efficient characteristics. It is difficult to efficiently train deep SNNs due to the non-differentiability of its activation function, which disables the typically used gradient descent approaches for traditional artificial neural networks (ANNs). Although the adoption of surrogate gradient (SG) formally allows for the back-propagation of losses, the discrete spiking mechanism actually differentiates the loss landscape of SNNs from that of ANNs, failing the surrogate gradient methods to achieve comparable accuracy as for ANNs. In this paper, we first analyze why the current direct training approach with surrogate gradient results in SNNs with poor generalizability. Then we introduce the temporal efficient training (TET) approach to compensate for the loss of momentum in the gradient descent with SG so that the training process can converge into flatter minima with better generalizability. Meanwhile, we demonstrate that TET improves the temporal scalability of SNN and induces a temporal inheritable training for acceleration. Our method consistently outperforms the SOTA on all reported mainstream datasets, including CIFAR-10/100 and ImageNet. Remarkably on DVS-CIFAR10, we obtained 83% top-1 accuracy, over 10% improvement compared to existing state of the art.

****

Dissecting Local Properties of Adversarial Examples
Lu Chen,Renjie Chen,Hang Guo,Yuan Luo,Quanshi Zhang,Yisen Wang

Adversarial examples have attracted significant attention over the years, yet a sufficient understanding is in lack, especially when analyzing their performances in combination with adversarial training. In this paper, we revisit some properties of adversarial examples from both frequency and spatial perspectives: 1) t he special high-frequency components of adversarial examples tend to mislead naturally-trained models while have little impact on adversarially-trained ones, and 2) adversarial examples show disorderly perturbations on naturally-trained models and locally-consistent (image shape related) perturbations on adversarially-trained ones. Motivated by these, we analyze the fragile tendency of models with the generated adversarial perturbations, and propose a connection with model vulnerability and local intermediate response. That is, a smaller local intermediate response comes along with better model adversarial robustness. To be specific, we demonstrate that: 1) DNNs are naturally fragile at least for large enough local response differences between adversarial/natural examples, 2) and smoother adversarially-trained models can alleviate local response differences with enhanced robustness.

****

Video Forgery Detection Using Multiple Cues on Fusion of EfficientNet and Swin Transformer

Chenyu Liu,Jia Li,Junxian Duan,Huaibo Huang

The rapid development of video processing technology makes it easy for people to forge videos without leaving visual artifacts. The spread of forged videos may lead to moral and legal consequences and pose a potential threat to people's lives and social stability. So it is significant to identify deepfake video information. Although the previous detection methods have achieved high accuracy, the generalization is poor when facing unprecedented data in the real scene. There are three fundamental reasons. The first is that capturing the general clue of artifacts is difficult. The second is that selecting the appropriate model is challenging in specific feature extraction. The third is that exploiting fully and effectively the extracted features is hard. We find that the high-frequency information in the image and the texture in the shallow layer of the model expose the subtle artifacts. The optical flow of the real video has variations while the optical flow of the deepfake video has rarely variations. Furthermore, consecutive frames in the real video have temporal consistency. In this paper, we propose a dual-branch video forgery detection model named ENST, which integrates parallelly and interactively EfficientNet-B5 and Swin Transformer. Specifically, EfficientNet-B5 extracts the artifacts information of high frequency and texture in the shallow layer of the model. Swin Transformer captures the subtle discrepancies between optical flows. To extract more robust face features, we design a new loss function for EfficientNet-B5. In addition, we also introduce the attention mechanism into EfficientNet-B5 to enhance the extracted features. We conduct test experiments on FaceForensics++ and Celeb-DF (v2) datasets, and comprehensive results show that ENST has higher accuracy and generalization, which is superior to the most advanced methods.

****

Reliable Adversarial Distillation with Unreliable Teachers

Jianing Zhu,Jiangchao Yao,Bo Han,Jingfeng Zhang,Tongliang Liu,Gang Niu,Jingren Zhou,Jianliang Xu,Hongxia Yang

In ordinary distillation, student networks are trained with soft labels (SLs) given by pretrained teacher networks, and students are expected to improve upon teachers since SLs are stronger supervision than the original hard labels. However, when considering adversarial robustness, teachers may become unreliable and adversarial distillation may not work: teachers are pretrained on their own adversarial data, and it is too demanding to require that teachers are also good at every adversarial data queried by students. Therefore, in this paper, we propose reliable introspective adversarial distillation (IAD) where students partially instead of fully trust their teachers. Specifically, IAD distinguishes between three cases given a query of a natural data (ND) and the corresponding adversarial data (AD): (a) if a teacher is good at AD, its SL is fully trusted; (b) if a teacher is good at ND but not AD, its SL is partially trusted and the student also takes its own SL into account; (c) otherwise, the student only relies on its own SL. Experiments demonstrate the effectiveness of IAD for improving upon teachers in terms of adversarial robustness.

****

Understanding the Interaction of Adversarial Training with Noisy Labels

Jianing Zhu,Jingfeng Zhang,Bo Han,Tongliang Liu,Gang Niu,Hongxia Yang,Mohan Kankanhalli,Masashi Sugiyama

Noisy labels (NL) and adversarial examples both undermine trained models, but interestingly they have hitherto been studied independently. A recent adversarial training (AT) study showed that the number of projected gradient descent (PGD) steps to successfully attack a point (i.e., find an adversarial example in its proximity) is an effective measure of the robustness of this point. Given that natural data are clean, this measure reveals an intrinsic geometric property---how far a point is from its nearest class boundary. Based on this breakthrough, in this paper, we figure out how AT would interact with NL. Firstly, we find if a point is too close to its noisy-class boundary (e.g., one step is enough to attack it), this point is likely to be mislabeled, which suggests to adopt the number of PGD steps as a new criterion for sample selection to correct NL. Secondly, we confirm that AT with strong smoothing effects suffers less from NL (without NL corrections) than standard training, which suggests that AT itself is an NL correction. Hence, AT with NL is helpful for improving even the natural accuracy, which again illustrates the superiority of AT as a general-purpose robust learning criterion.

****

Neural Program Synthesis with Query

Di Huang,Rui Zhang,Xing Hu,Xishan Zhang,Pengwei Jin,Nan Li,Zidong Du,Qi Guo,Yunji Chen

Aiming to find a program satisfying the user intent given input-output examples, program synthesis has attracted increasing interest in the area of machine learning. Despite the promising performance of existing methods, most of their success comes from the privileged information of well-designed input-output examples. However, providing such input-output examples is unrealistic because it requires the users to have the ability to describe the underlying program with a few input-output examples under the training distribution. In this work, we propose a query-based framework that trains a query neural network to generate informative input-output examples automatically and interactively from a large query space. The quality of the query depends on the amount of the mutual information between the query and the corresponding program, which can guide the optimization of the query framework. To estimate the mutual information more accurately, we introduce the functional space (F-space) which models the relevance between the input-output examples and the programs in a differentiable way. We evaluate the effectiveness and generalization of the proposed query-based framework on the Karel task and the list processing task. Experimental results show that the query-based framework can generate informative input-output examples which achieve and even outperform well-designed input-output examples.

****

Delaunay Component Analysis for Evaluation of Data Representations

Petra Poklukar,Vladislav Polianskii,Anastasiia Varava,Florian T. Pokorny,Danica Kragic Jensfelt

Advanced representation learning techniques require reliable and general evaluation methods. Recently, several algorithms based on the common idea of geometric and topological analysis of a manifold approximated from the learned data representations have been proposed. In this work, we introduce Delaunay Component Analysis (DCA) -- an evaluation algorithm which approximates the data manifold using a more suitable neighbourhood graph called Delaunay graph. This provides a reliable manifold estimation even for challenging geometric arrangements of representations such as clusters with varying shape and density as well as outliers, which is where existing methods often fail. Furthermore, we exploit the nature of Delaunay graphs and introduce a framework for assessing the quality of individual novel data representations. We experimentally validate the proposed DCA method on representations obtained from neural networks trained with contrastive objective, supervised and generative models, and demonstrate various use cases of our extended single point evaluation framework.

****

Abelian Neural Networks

Kenshin Abe,Takanori Maehara,Issei Sato

In several domains such as natural language processing, it has been empirically reported that simple addition and subtraction in a somehow learned embedding space capture analogical relations. However, there is no guarantee that such relation holds for a new embedding space acquired by some training strategies. To tackle this issue, we propose to explicitly model analogical structure with an Abelian group. We construct an Abelian group network using invertible neural networks and show its universal approximation property. In experiments, our model successfully learns to capture word analogies from word2vec representations and shows better performance than other learning-based strategies. As a byproduct of modeling Abelian group operations, we furthermore obtain its natural extension to permutation invariant models with theoretical size-generalization capability.

****

Visual hyperacuity with moving sensor and recurrent neural computations

Alexander Rivkind,Or Ram,Eldad Assa,Michael Kreiserman,Ehud Ahissar

Dynamical phenomena, such as recurrent neuronal activity and perpetual motion of the eye, are typically overlooked in models of bottom-up visual perception. Recent experiments suggest that tiny inter-saccadic eye motion ("fixational drift") enhances visual acuity beyond the limit imposed by the density of retinal photoreceptors. Here we hypothesize that such an enhancement is enabled by recurrent neuronal computations in early visual areas. Specifically, we explore a setting involving a low-resolution dynamical sensor that moves with respect to a static scene, with drift-like tiny steps. This setting mimics a dynamical eye viewing objects in perceptually-challenging conditions. The dynamical sensory input is classified by a convolutional neural network with recurrent connectivity added to its lower layers, in analogy to recurrent connectivity in early visual areas. Applying our system to CIFAR-10 and CIFAR-100 datasets down-sampled via 8x8 sensor, we found that (i) classification accuracy, which is drastically reduced by this down-sampling, is mostly restored to its 32x32 baseline level when using a moving sensor and recurrent connectivity, (ii) in this setting, neurons in the early layers exhibit a wide repertoire of selectivity patterns, spanning the spatiotemporal selectivity space, with neurons preferring different combinations of spatial and temporal patterning, and (iii) curved sensor's trajectories improve visual acuity compared to straight trajectories, echoing recent experimental findings involving eye-tracking in challenging conditions. Our work sheds light on the possible role of recurrent connectivity in early vision as well as the roles of fixational drift and temporal-frequency selective cells in the visual system. It also proposes a solution for artificial image recognition in settings with limited resolution and multiple time samples, such as in edge AI applications.

****

Neural Manifold Clustering and Embedding

ZENGYI LI,Yubei Chen,Yann LeCun,Friedrich Sommer

Given a union of non-linear manifolds, non-linear subspace clustering or manifold clustering aims to cluster data points based on manifold structures and also learn to parameterize each manifold as a linear subspace in a feature space. Deep neural networks have the potential to achieve this goal under highly non-linear settings given their large capacity and flexibility. We argue that achieving manifold clustering with neural networks requires two essential ingredients: a domain-specific constraint that ensures the identification of the manifolds, and a learning algorithm for embedding each manifold to a linear subspace in the feature space. This work shows that many constraints can be implemented by data augmentation. For subspace feature learning, Maximum Coding Rate Reduction (MCR^2) objective can be used. Putting them together yields Neural Manifold Clustering and Embedding (NMCE), a novel method for general purpose manifold clustering, which significantly outperforms autoencoder-based deep subspace clustering and achieve state-of-the-art performance on several important benchmarks. Further, on more challenging natural image datasets, NMCE can also outperform other algorithms specifically designed for clustering. Qualitatively, we demonstrate that NMCE learns a meaningful and interpretable feature space. As the formulation of NMCE i s closely related to several important Self-supervised learning (SSL) methods, we believe this work can help us build a deep understanding on SSL representation learning.

****

Fully Decentralized Model-based Policy Optimization with Networked Agents

Yuchen Liu,Yali Du,Runji Lin,Hangrui Bi,Mingdong Wu,Jun Wang,Hao Dong

Model-based RL is an effective approach for reducing sample complexity. However, when it comes to multi-agent setting where the number of agent is large, the model estimation can be problematic due to the exponential increased interactions. In this paper, we propose a decentralized model-based reinforcement learning algorithm for networked multi-agent systems, where agents are cooperative and communicate locally with their neighbors. We analyze our algorithm theoretically and derive an upper bound of performance discrepancy caused by model usage, and provide a sufficient condition of monotonic policy improvement. In our experiments, we compare our algorithm against other strong multi-agent baselines and demonstrate that our algorithm not only matches the asymptotic performance of model-free methods but also largely increases its sample efficiency.

****

Partial Wasserstein Adversarial Network for Non-rigid Point Set Registration

Ziming Wang,Nan Xue,Ling Lei,Gui-Song Xia

Given two point sets, the problem of registration is to recover a transformation that matches one set to the other. This task is challenging due to the presence of large number of outliers, the unknown non-rigid deformations and the large sizes of point sets. To obtain strong robustness against outliers, we formulate the registration problem as a partial distribution matching (PDM) problem, where the goal is to partially match the distributions represented by point sets in a metric space. To handle large point sets, we propose a scalable PDM algorithm by utilizing the efficient partial Wasserstein-1 (PW) discrepancy. Specifically, we derive the Kantorovich-Rubinstein duality for the PW discrepancy, and show its gradient can be explicitly computed. Based on these results, we propose a partial Wasserstein adversarial network (PWAN), which is able to approximate the PW discrepancy by a neural network, and minimize it by gradient descent. In addition,
it also incorporates an efficient coherence regularizer for non-rigid transformations to avoid unrealistic deformations. We evaluate PWAN on practical point set registration tasks, and show that the proposed PWAN is robust, scalable and performs more favorably than the state-of-the-art methods.

****

Finding lost DG: Explaining domain generalization via model complexity

Da Li,Henry Gouk,Timothy Hospedales

The domain generalization (DG) problem setting challenges a model trained on multiple known data distributions to generalise well on unseen data distributions. Due to its practical importance, a large number of methods have been proposed to address this challenge. However most of this work is empirical, as the DG problem is hard to model formally; and recent evaluations have cast doubt on existing methods’ practical efficacy -- in particular compared to a well chosen empirical risk minimisation baseline.

We present a novel learning-theoretic generalisation bound for DG that bounds novel domain generalisation performance in terms of the model’s Rademacher complexity. Based on this, we conjecture that the causal factor behind existing methods’ efficacy or lack thereof is a variant of the standard empirical risk-predictor complexity tradeoff, and demonstrate that their performance variability can be explained in these terms. Algorithmically, this analysis suggests that domain generalisation should be achieved by simply performing regularised ERM with a leave-one-domain-out cross-validation objective. Empirical results on the DomainBed benchmark corroborate this.

****

Best Practices in Pool-based Active Learning for Image Classification

Adrian Lang,Christoph Mayer,Radu Timofte

The recent popularity of active learning (AL) methods for image classification using deep-learning has led to a large number of publications that lead to significant progress in the field. Benchmarking the latest works in an exhaustive and unified way and evaluating the improvements made by the novel methods is of key importance to advance the research in AL. Reproducing state-of-the-art AL methods is often cumbersome, since the results and the ranking order of different strategies are highly dependent on several factors, such as training settings, used data type, network architectures, loss function and more. With our work we highlight the main factors that should be considered when proposing new AL strategies. In addition, we provide solid benchmarks to compare new with existing methods. We therefore conduct a comprehensive study on the influence of these key aspects, providing best practices in pool-based AL for image classification. We emphasize aspects such as the importance of using data augmentation, the need of separating the contribution of a classification network and the acquisition strategy to the overall performance, the advantages that a proper initialization of the network can bring to AL. Moreover, we make a new codebase available, that enables state-of-the-art performance for the investigated methods, which we hope will serve the AL community as a new starting point when proposing new AL strategies.

****

Generalized Decision Transformer for Offline Hindsight Information Matching

Hiroki Furuta,Yutaka Matsuo,Shixiang Shane Gu

How to extract as much learning signal from each trajectory data has been a key problem in reinforcement learning (RL), where sample inefficiency has posed serious challenges for practical applications. Recent works have shown that using expressive policy function approximators and conditioning on future trajectory information -- such as future states in hindsight experience replay (HER) or returns-to-go in Decision Transformer (DT) -- enables efficient learning of multi-task policies, where at times online RL is fully replaced by offline behavioral cloning (BC), e.g. sequence modeling. We demonstrate that all these approaches are doing hindsight information matching (HIM) -- training policies that can output the rest of trajectory that matches some statistics of future state information. We present Generalized Decision Transformer (GDT) for solving any HIM problem, and show how different choices for the feature function and the anti-causal aggregator not only recover DT as a special case, but also lead to novel Categorical DT (CDT) and Bi-directional DT (BDT) for matching different statistics of the future. For evaluating CDT and BDT, we define offline multi-task state-marginal matching (SMM) and imitation learning (IL) as two generic HIM problems, propose a Wasserstein distance loss as a metric for both, and empirically study them on MuJoCo continuous control benchmarks. Categorical DT, which simply replaces anti-causal summation with anti-causal binning in DT, enables arguably the first effective offline multi-task SMM algorithm that generalizes well to unseen (and even synthetic) multi-modal reward or state-feature distributions. Bi-directional DT, which uses an anti-causal second transformer as the aggregator, can learn to model any statistics of the future and outperforms DT variants in offline multi-task IL, i.e. one-shot IL. Our generalized formulations from HIM and GDT greatly expand the role of powerful sequence modeling architectures in modern RL.

****

Towards Unsupervised Content Disentanglement in Sentence Representations via Syntactic Roles

Ghazi Felhi,Joseph Le Roux,Djamé Seddah

Linking neural representations to linguistic factors is crucial in order to build and analyze NLP models interpretable by humans. Among these factors, syntactic roles (e.g. subjects, direct objects,$\dots$) and their realizations are essential markers since they can be understood as a decomposition of predicative structures and thus the meaning of sentences. Starting from a deep probabilistic generative model with attention, we measure the interaction between latent variables and realizations of syntactic roles, and show that it is possible to obtain, without supervision, representations of sentences where different syntactic roles correspond to clearly identified different latent variables. The probabilistic model we propose is an Attention-Driven Variational Autoencoder (ADVAE). Drawing inspiration from Transformer-based machine translation models, ADVAEs enable the analysis of the interactions between latent variables and input tokens through attention. We also develop an evaluation protocol to measure disentanglement with regard to the realizations of syntactic roles. This protocol is based on attention maxima for the encoder and on disturbing individual latent variables for the decoder. Our experiments on raw English text from the SNLI dataset show that $\textit{i)}$ disentanglement of syntactic roles can be induced without supervision, $\textit{ii)}$ ADVAE separates more syntactic roles than classical sequence VAEs, $\textit{iii)}$ realizations of syntactic roles can be separately modified in sentences by mere intervention on the associated latent variables. Our work constitutes a first step towards unsupervised controllable content generation. The code for our work is publicly available.

****

Analytically Tractable Bayesian Deep Q-Learning

Luong-Ha Nguyen,James-A. Goulet

Reinforcement learning (RL) has gained increasing interest since the demonstration it was able to reach human performance on video game benchmarks using deep Q-learning (DQN). The current consensus of DQN for training neural networks (NNs) on such complex environments is to rely on gradient-descent optimization (GD). This consensus ignores the uncertainty of the NN's parameters which is a key aspect for the selection of an optimal action given a state. Although alternative Bayesian deep learning methods exist, most of them still rely on GD and numerical approximations, and they typically do not scale on complex benchmarks such as the Atari game environment. In this paper, we present how we can adapt the temporal difference Q-learning framework to make it compatible with the tractable approximate Gaussian inference (TAGI) which allows estimating the posterior distribution of NN's parameters using a closed-form analytical method. Throughout the experiments with on- and off-policy reinforcement learning approaches, we demonstrate that TAGI can reach a performance comparable to backpropagation-trained networks while using only half the number of hyperparameters, and without relying on GD or numerical approximations.

****

Quantitative Performance Assessment of CNN Units via Topological Entropy Calculation

Yang Zhao,Hao Zhang

Identifying the status of individual network units is critical for understanding the mechanism of convolutional neural networks (CNNs). However, it is still challenging to reliably give a general indication of unit status, especially for units in different network models. To this end, we propose a novel method for quantitatively clarifying the status of single unit in CNN using algebraic topological tools. Unit status is indicated via the calculation of a defined topological-based entropy, called feature entropy, which measures the degree of chaos of the global spatial pattern hidden in the unit for a category. In this way, feature entropy could provide an accurate indication of status for units in different networks with diverse situations like weight-rescaling operation. Further, we show that feature entropy decreases as the layer goes deeper and shares almost simultaneous trend with loss during training. We show that by investigating the feature entropy of units on only training data, it could give discrimination between networks with different generalization ability from the view of the effectiveness of feature representations.

****

Learning Neural Causal Models with Active Interventions

Nino Scherrer,Olexa Bilaniuk,Yashas Annadani,Anirudh Goyal,Patrick Schwab,Bernhard Schölkopf,Michael Curtis Mozer,Yoshua Bengio,Stefan Bauer,Nan Rosemary Ke

Discovering causal structures from data is a challenging inference problem of fundamental importance in all areas of science. The appealing scaling properties of neural networks have recently led to a surge of interest in differentiable neural network-based methods for learning causal structures from data. So far, differentiable causal discovery has focused on static datasets of observational or i

nterventional origin. In this work, we introduce an active intervention-targeting mechanism which enables quick identification of the underlying causal structure of the data-generating process. Our method significantly reduces the required number of interactions compared with random intervention targeting and is applicable for both discrete and continuous optimization formulations of learning the underlying directed acyclic graph (DAG) from data. We examine the proposed method across multiple frameworks in a wide range of settings and demonstrate superior performance on multiple benchmarks from simulated to real-world data.
****

Adaptive Inertia: Disentangling the Effects of Adaptive Learning Rate and Momentum

Zeke Xie,Xinrui Wang,Huishuai Zhang,Issei Sato,Masashi Sugiyama

Adaptive Momentum Estimation (Adam), which combines Adaptive Learning Rate and Momentum, would be the most popular stochastic optimizer for accelerating the training of deep neural networks. However, empirically Adam often generalizes worse than Stochastic Gradient Descent (SGD). We unveil the mystery of this behavior in the diffusion theoretical framework. Specifically, we disentangle the effects of Adaptive Learning Rate and Momentum of the Adam dynamics on saddle-point escaping and flat minima selection. We prove that Adaptive Learning Rate can escape saddle points efficiently, but cannot select flat minima as SGD does. In contrast, Momentum provides a drift effect to help the training process pass through saddle points, and almost does not affect flat minima selection. This partly explains why SGD (with Momentum) generalizes better, while Adam generalizes worse but converges faster. Furthermore, motivated by the analysis, we design a novel adaptive optimization framework named Adaptive Inertia, which uses parameter-wise adaptive inertia to accelerate the training and provably favors flat minima as well as SGD. Our extensive experiments demonstrate that the proposed adaptive inertia method can generalize significantly better than SGD and conventional adaptive gradient methods.
****

Contrastively Enforcing Distinctiveness for Multi-Label Classification

Son Duy Dao,He Zhao,Dinh Phung,Jianfei Cai

Recently, as an effective way of learning latent representations, contrastive learning has been increasingly popular and successful in various domains. The success of constrastive learning in single-label classifications motivates us to leverage this learning framework to enhance distinctiveness for better performance in multi-label image classification. In this paper, we show that a direct application of contrastive learning can hardly improve in multi-label cases. Accordingly, we propose a novel framework for multi-label classification with contrastive learning in a fully supervised setting, which learns multiple representations of an image under the context of different labels. This facilities a simple yet intuitive adaption of contrastive learning into our model to boost its performance in multi-label image classification.

Extensive experiments on two benchmark datasets show that the proposed framework achieves state-of-the-art performance in the comparison with the advanced methods in multi-label classification.
****

Imitation Learning by Reinforcement Learning

Kamil Ciosek

Imitation learning algorithms learn a policy from demonstrations of expert behavior. We show that, for deterministic experts, imitation learning can be done by reduction to reinforcement learning with a stationary reward. Our theoretical analysis both certifies the recovery of expert reward and bounds the total variation distance between the expert and the imitation learner, showing a link to adversarial imitation learning. We conduct experiments which confirm that our reduction works well in practice for continuous control tasks.
****

Non-Denoising Forward-Time Diffusions

Stefano Peluchetti

The scope of this paper is generative modeling through diffusion processes.

An approach falling within this paradigm is the work of Song et al. (2021), which relies on a time-reversal argument to construct a diffusion process targeting the desired data distribution.

We show that the time-reversal argument, common to all denoising diffusion probabilistic modeling proposals, is not necessary.

We obtain diffusion processes targeting the desired data distribution by taking appropriate mixtures of diffusion bridges.

The resulting transport is exact by construction, allows for greater flexibility in choosing the dynamics of the underlying diffusion, and can be approximated by means of a neural network via novel training objectives.

We develop an unifying view of the drift adjustments corresponding to our and to time-reversal approaches and make use of this representation to inspect the inner workings of diffusion-based generative models.

Finally, we leverage on scalable simulation and inference techniques common in spatial statistics to move beyond fully factorial distributions in the underlying diffusion dynamics.

The methodological advances contained in this work contribute toward establishing a general framework for generative modeling based on diffusion processes.

****

On-Policy Model Errors in Reinforcement Learning

Lukas Froehlich,Maksym Lefarov,Melanie Zeilinger,Felix Berkenkamp

Model-free reinforcement learning algorithms can compute policy gradients given sampled environment transitions, but require large amounts of data. In contrast, model-based methods can use the learned model to generate new data, but model errors and bias can render learning unstable or suboptimal. In this paper, we present a novel method that combines real-world data and a learned model in order to get the best of both worlds. The core idea is to exploit the real-world data for on-policy predictions and use the learned model only to generalize to different actions. Specifically, we use the data as time-dependent on-policy correction terms on top of a learned model, to retain the ability to generate data without accumulating errors over long prediction horizons. We motivate this method theoretically and show that it counteracts an error term for model-based policy improvement. Experiments on MuJoCo- and PyBullet-benchmarks show that our method can drastically improve existing model-based approaches without introducing additional tuning parameters.

****

Neuronal Learning Analysis using Cycle-Consistent Adversarial Networks

Bryan M. Li,Theoklitos Amvrosiadis,Nathalie L. Rochefort,Arno Onken

Understanding how activity in neural circuits reshapes following task learning could reveal fundamental mechanisms of learning. Thanks to the recent advances in neural imaging technologies, high-quality recordings can be obtained from hundreds of neurons over multiple days or even weeks. However, the complexity and dimensionality of population responses pose significant challenges for analysis. Existing methods of studying neuronal adaptation and learning often impose strong assumptions on the data or model, resulting in biased descriptions that do not generalize. In this work, we use a variant of deep generative models called - cycle-consistent adversarial networks, to learn the unknown mapping between pre- and post-learning neuronal activities recorded $\textit{in vivo}$. To do so, we develop an end-to-end pipeline to preprocess, train and evaluate calcium fluorescence signals, and a procedure to interpret the resulting deep learning models. To assess the validity of our method, we first test our framework on a synthetic dataset with known ground-truth transformation. Subsequently, we applied our method to neuronal activities recorded from the primary visual cortex of behaving mice, where the mice transition from novice to expert-level performance in a visual-based virtual reality experiment. We evaluate model performance on generated calcium imaging signals and their inferred spike trains. To maximize performance, we derive a novel approach to pre-sort neurons such that convolutional-based networks can take advantage of the spatial information that exists in neuronal activities. In addition, we incorporate visual explanation methods to improve the interpretability of our work and gain insights into the learning process as manif ested in the cellular activities. Together, our results demonstrate that analyzing neuronal learning processes with data-driven deep unsupervised methods holds the potential to unravel changes in an unbiased way.

****

TAPEX: Table Pre-training via Learning a Neural SQL Executor

Qian Liu,Bei Chen,Jiaqi Guo,Morteza Ziyadi,Zeqi Lin,Weizhu Chen,Jian-Guang Lou

Recent progress in language model pre-training has achieved a great success via leveraging large-scale unstructured textual data. However, it is still a challenge to apply pre-training on structured tabular data due to the absence of large-scale high-quality tabular data. In this paper, we propose TAPEX to show that table pre-training can be achieved by learning a neural SQL executor over a synthetic corpus, which is obtained by automatically synthesizing executable SQL queries and their execution outputs. TAPEX addresses the data scarcity challenge via guiding the language model to mimic a SQL executor on the diverse, large-scale and high-quality synthetic corpus. We evaluate TAPEX on four benchmark datasets. Experimental results demonstrate that TAPEX outperforms previous table pre-training approaches by a large margin and achieves new state-of-the-art results on all of them. This includes the improvements on the weakly-supervised WikiSQL denotation accuracy to 89.5% (+2.3%), the WikiTableQuestions denotation accuracy to 57.5% (+4.8%), the SQA denotation accuracy to 74.5% (+3.5%), and the TabFact accuracy to 84.2% (+3.2%). To our knowledge, this is the first work to exploit table pre-training via synthetic executable programs and to achieve new state-of-the-art results on various downstream tasks. Our code can be found at https://github.com/microsoft/Table-Pretraining.

****

Succinct Compression: Near-Optimal and Lossless Compression of Deep Neural Networks during Inference Runtime

Yicun Duan,Xiangjun Peng

Recent advances in Deep Neural Networks (DNN) compression (e.g. pruning, quantization and etc.) significantly reduces the amount of space consumption for storage, making them easier to deploy in low-cost devices. However, those techniques do not keep the compressed representation during inference runtime, which incurs significant overheads in terms of both performance and space consumption. We introduce ``Succinct Compression”, a three-stage framework to enable DNN inference with near-optimal compression and much better performance during inference runtime. The key insight of our method leverages the concept of \textit{Succinct Data Structures}, which supports fast queries directly on compressed representation without decompression. Our method first transforms DNN models as our proposed formulations in either Element-wise or Block-wise manner, so that \textit{Succinct Data Structures} can take advantage of. Then, our method compresses transformed DNN models using \textit{Succinct Data Structures}. Finally, our method exploits our specialized execution pipelines for different model formulations, to retrieve relevant data for DNN inference. Our experimental results show that, our method keeps near-optimal compression, and achieves at least 8.7X/11.5X speedup on AlexNet/VGG-16 inference, compared with Huffman Coding. We also experimentally show that our method is quite synergistic with Pruning and Quantization.

****

Transfer and Marginalize: Explaining Away Label Noise with Privileged Information

Mark Collier,Rodolphe Jenatton,Effrosyni Kokiopoulou,Jesse Berent

Supervised learning datasets often have privileged information, in the form of features which are available at training time but are not available at test time e.g. the ID of the annotator that provided the label. We argue that privileged information is useful for explaining away label noise, thereby reducing the harmful impact of noisy labels. We develop a simple and efficient method for supervised neural networks: it transfers the knowledge learned with privileged information via weight sharing and approximately marginalizes over privileged information at test time. Our method, TRAM (TRansfer and Marginalize), has minimal training time overhead and has the same test time cost as not using privileged informati

on. TRAM performs strongly on CIFAR-10H, ImageNet and Civil Comments benchmarks.
****

Two Regimes of Generalization for Non-Linear Metric Learning
Mark Kozdoba,Shie Mannor

A common approach to metric learning is to seek an embedding of the input data that behaves well with respect to the labels. While generalization bounds for linear embeddings are known, the non-linear case is not well understood. In this work we fill this gap by providing uniform generalization guarantees for the case where the metric is induced by a neural network type embedding of the data. Specifically, we discover and analyze two regimes of behavior of the networks, which are roughly related to the sparsity of the last layer. The bounds corresponding to the first regime are based on the spectral and $(2,1)$-norms of the weight matrices, while the second regime bounds use the $(2,\infty)$-norm at the last layer, and are significantly stronger when the last layer is dense. In addition, we empirically evaluate the behavior of the bounds for networks trained with SGD on the MNIST and 20newsgroups datasets. In particular, we demonstrate that both regimes occur naturally on realistic data.
****

Graph Convolutional Networks via Adaptive Filter Banks
Xing Gao,Wenrui Dai,Chenglin Li,Junni Zou,Hongkai Xiong,Pascal Frossard

Graph convolutional networks have been a powerful tool in representation learning of networked data. However, most architectures of message passing graph convolutional networks (MPGCNs) are limited as they employ a single message passing strategy and typically focus on low-frequency information, especially when graph features or signals are heterogeneous in different dimensions. Then, existing spectral graph convolutional operators lack a proper sharing scheme between filters, which may result in overfitting problems with numerous parameters. In this paper, we present a novel graph convolution operator, termed BankGCN, which extends the capabilities of MPGCNs beyond single `low-pass' features and simplifies spectral methods with a carefully designed sharing scheme between filters. BankGCN decomposes multi-channel signals on arbitrary graphs into subspaces and shares adaptive filters to represent information in each subspace. The filters of all subspaces differ in frequency response and together form a filter bank. The filter bank and the signal decomposition permit to adaptively capture diverse spectral characteristics of graph data for target applications with a compact architecture. We finally show through extensive experiments that BankGCN achieves excellent performance on a collection of benchmark graph datasets.
****

Unsupervised Pose-Aware Part Decomposition for 3D Articulated Objects
Yuki Kawana,YUSUKE Mukuta,Tatsuya Harada

Articulated objects exist widely in the real world. However, previous 3D generative methods for unsupervised part decomposition are unsuitable for such objects, because they assume a spatially fixed part location, resulting in inconsistent part parsing. In this paper, we propose PPD (unsupervised Pose-aware Part Decomposition) to address a novel setting that explicitly targets man-made articulated objects with mechanical joints, considering the part poses. We show that category-common prior learning for both part shapes and poses facilitates the unsupervised learning of (1) part decomposition with non-primitive-based implicit representation, and (2) part pose as joint parameters under single-frame shape supervision. We evaluate our method on synthetic and real datasets, and we show that it outperforms previous works in consistent part parsing of the articulated objects based on comparable part pose estimation performance to the supervised baseline.
****

DARA: Dynamics-Aware Reward Augmentation in Offline Reinforcement Learning
Jinxin Liu,Zhang Hongyin,Donglin Wang

Offline reinforcement learning algorithms promise to be applicable in settings where a fixed dataset is available and no new experience can be acquired. However, such formulation is inevitably offline-data-hungry and, in practice, collectin

g a large offline dataset for one specific task over one specific environment is also costly and laborious. In this paper, we thus 1) formulate the offline dynamics adaptation by using (source) offline data collected from another dynamics to relax the requirement for the extensive (target) offline data, 2) characterize the dynamics shift problem in which prior offline methods do not scale well, and 3) derive a simple dynamics-aware reward augmentation (DARA) framework from both model-free and model-based offline settings. Specifically, DARA emphasizes learning from those source transition pairs that are adaptive for the target environment and mitigates the offline dynamics shift by characterizing state-action-next-state pairs instead of the typical state-action distribution sketched by prior offline RL methods. The experimental evaluation demonstrates that DARA, by augmenting rewards in the source offline dataset, can acquire an adaptive policy for the target environment and yet significantly reduce the requirement of target offline data. With only modest amounts of target offline data, our performance consistently outperforms the prior offline RL methods in both simulated and real-world tasks.
****

L2BGAN: An image enhancement model for image quality improvement and image analysis tasks without paired supervision
Jhilik Bhattacharya,Gianni Ramponi,Leonardo Gregorat,Shatrughan Modi
The paper presents an image enhancement model, L2BGAN, to translate low light images to bright images without a paired supervision. We introduce the use of geometric and lighting consistency along with a contextual loss criterion. These when combined with multiscale color, texture and edge discriminators prove to provide competitive results. We perform extensive experiments on benchmark datasets to compare our results visually as well as objectively. We observe the performance of L2BGAN on real time driving datasets which are subject to motion blur, noise and other artifacts. We further demonstrate the application of image understanding tasks on our enhanced images using DarkFace and ExDark datasets.

****

Towards Structured Dynamic Sparse Pre-Training of BERT
Anastasia S. D. Dietrich,Frithjof Gressmann,Douglas Orr,Ivan Chelombiev,Daniel Justus,Carlo Luschi
Identifying algorithms for computational efficient unsupervised training of large language models is an important and active area of research.
In this work, we develop and study a straightforward, dynamic always-sparse pre-training approach for BERT language modeling, which leverages periodic compression steps based on magnitude pruning followed by random parameter re-allocation.
This approach enables us to achieve Pareto improvements in terms of the number of floating-point operations (FLOPs) over statically sparse and dense models across a broad spectrum of network sizes.
Furthermore, we demonstrate that training remains FLOP-efficient when using coarse-grained block sparsity, making it particularly promising for efficient execution on modern hardware accelerators.
****

Understanding Knowledge Integration in Language Models with Graph Convolutions
Yifan Hou,Guoji Fu,MRINMAYA SACHAN
Pretrained language models (LMs) are not very good at robustly capturing factual knowledge. This has led to the development of a number of knowledge integration (KI) methods which aim to incorporate external knowledge into pretrained LMs. Even though KI methods show some performance gains over base LMs, the efficacy and limitations of these methods are not well-understood. For instance, it is unclear how and what kind of knowledge is effectively integrated into LMs and if such integration may lead to catastrophic forgetting of already learned knowledge.
In this paper, we revisit the KI process from the view of graph signal processin

g and show that KI could be interpreted using a graph convolution operation. We propose a simple probe model called Graph Convolution Simulator (GCS) for interpreting knowledge-enhanced LMs and exposing what kind of knowledge is integrated into these models. We conduct experiments to verify that our GCS model can indeed be used to correctly interpret the KI process, and we use it to analyze two typical knowledge-enhanced LMs: K-Adapter and ERNIE. We find that only a small amount of factual knowledge is captured in these models during integration. While K-Adapter is better at integrating simple relational knowledge, complex relational knowledge is integrated better in ERNIE. We further find that while K-Adapter struggles to integrate time-related knowledge, it successfully integrates knowledge of unpopular entities and relations. Our analysis also show some challenges in KI. In particular, we find simply increasing the size of the KI corpus may not lead to better KI and more fundamental advances may be needed.
****

The Manifold Hypothesis for Gradient-Based Explanations
Sebastian Bordt,Uddeshya Upadhyay,Zeynep Akata,Ulrike von Luxburg
When are gradient-based explanations meaningful? We propose a necessary criterion: explanations need to be aligned with the tangent space of the data manifold. To test this hypothesis, we employ autoencoders to estimate and generate data manifolds. Across a range of different datasets -- MNIST, EMNIST, CIFAR10, X-ray pneumonia and Diabetic Retinopathy detection -- we demonstrate empirically that the more an explanation is aligned with the tangent space of the data, the more interpretable it tends to be. In particular, popular post-hoc explanation methods such as Integrated Gradients and SmoothGrad tend to align their results with the data manifold. The same is true for the outcome of adversarial training, which has been claimed to lead to more interpretable explanations. Empirically, alignment with the data manifold happens early during training, and to some degree even when training with random labels. However, we theoretically prove that good generalization of neural networks does not imply good or bad alignment of model gradients with the data manifold. This leads to a number of interesting follow-up questions regarding gradient-based explanations.

****

Multivariate Time Series Forecasting with Latent Graph Inference
Victor Garcia Satorras,Syama Sundar Rangapuram,Tim Januschowski
This paper introduces a new architecture for multivariate time series forecasting that simultaneously infers and leverages relations among time series. We cast our method as a modular extension to univariate architectures where relations among individual time series are dynamically inferred in the latent space obtained after encoding the whole input signal. Our approach is flexible enough to scale gracefully according to the needs of the forecasting task under consideration. In its most straight-forward and general version, we infer a potentially fully connected graph to model the interactions between time series, which allows us to obtain competitive forecast accuracy compared with the state-of-the-art in graph neural networks for forecasting. In addition, whereas previous latent graph inference methods scale O(N^2) w.r.t. the number of nodes N (representing the time series), we show how to configure our approach to cater for the scale of modern time series panels. By assuming the inferred graph to be bipartite where one partition consists of the original N nodes and we introduce K nodes (taking inspiration from low-rank-decompositions) we reduce the time complexity of our procedure to O(NK). This allows us to leverage the dependency structure with a small trade-off in forecasting accuracy. We demonstrate the effectiveness of our method for a variety of datasets where it performs better or very competitively to previous methods under both the fully connected and bipartite assumptions.
****

Explaining Point Processes by Learning Interpretable Temporal Logic Rules
Shuang Li,Mingquan Feng,Lu Wang,Abdelmajid Essofi,Yufeng Cao,Junchi Yan,Le Song
We propose a principled method to learn a set of human-readable logic rules to explain temporal point processes.

We assume that the generative mechanisms underlying the temporal point processes are governed by a set of first-order temporal logic rules, as a compact representation of domain knowledge. Our method formulates the rule discovery process from noisy event data as a maximum likelihood problem, and designs an efficient and tractable branch-and-price algorithm to progressively search for new rules and expand existing rules. The proposed algorithm alternates between the rule generation stage and the rule evaluation stage, and uncovers the most important collection of logic rules within a fixed time limit for both synthetic and real event data. In a real healthcare application, we also had human experts (i.e., doctors) verify the learned temporal logic rules and provide further improvements. These expert-revised interpretable rules lead to a point process model which outperforms previous state-of-the-arts for symptom prediction, both in their occurrence times and types.

****

Unifying Likelihood-free Inference with Black-box Optimization and Beyond

Dinghuai Zhang,Jie Fu,Yoshua Bengio,Aaron Courville

Black-box optimization formulations for biological sequence design have drawn recent attention due to their promising potential impact on the pharmaceutical industry. In this work, we propose to unify two seemingly distinct worlds: likelihood-free inference and black-box optimization, under one probabilistic framework. In tandem, we provide a recipe for constructing various sequence design methods based on this framework. We show how previous optimization approaches can be "reinvented" in our framework, and further propose new probabilistic black-box optimization algorithms. Extensive experiments on sequence design application illustrate the benefits of the proposed methodology.

****

Exploring Non-Contrastive Representation Learning for Deep Clustering

Zhizhong Huang,Jie Chen,Junping Zhang,Hongming Shan

Existing deep clustering methods rely on contrastive learning for representation learning, which require negative examples to form an embedding space where all instances are well-separated. However, the negative examples inevitably give rise to the class collision issue, compromising the representation learning for clustering. In this paper, we explore the non-contrastive representation learning for deep clustering, termed NCC, which is based on BYOL, a representative method without negative examples. First, we propose a positive sampling strategy to align one augmented view of instance with the neighbors of another view so that we can avoid the class collision issue caused by the negative examples and hence improve the within-cluster compactness. Second, we propose a novel prototypical contrastive loss, ProtoCL, which can encourage prototypical alignment between two augmented views and prototypical uniformity, hence maximizing the inter-cluster distance. Moreover, we formulate NCC in an Expectation-Maximization (EM) framework, in which E-step utilizes spherical k-means to estimate the pseudo-labels of instances and distribution of prototypes from the target network and M-step leverages the proposed losses to optimize the online network. As a result, NCC is able to form an embedding space where all clusters are well-separated and within-cluster examples are compact. Experimental results on several clustering benchmark datasets as well as ImageNet-1K demonstrate that the proposed NCC outperforms the state-of-the-art methods by a significant margin.

****

DEPTS: Deep Expansion Learning for Periodic Time Series Forecasting

Wei Fan,Shun Zheng,Xiaohan Yi,Wei Cao,Yanjie Fu,Jiang Bian,Tie-Yan Liu

Periodic time series (PTS) forecasting plays a crucial role in a variety of industries to foster critical tasks, such as early warning, pre-planning, resource scheduling, etc. However, the complicated dependencies of the PTS signal on its inherent periodicity as well as the sophisticated composition of various periods hinder the performance of PTS forecasting. In this paper, we introduce a deep expansion learning framework, DEPTS, for PTS forecasting. DEPTS starts with a decoupled formulation by introducing the periodic state as a hidden variable, which stimulates us to make two dedicated modules to tackle the aforementioned two challenges. First, we develop an expansion module on top of residual learning to pe rform a layer-by-layer expansion of those complicated dependencies. Second, we introduce a periodicity module with a parameterized periodic function that holds sufficient capacity to capture diversified periods. Moreover, our two customized modules also have certain interpretable capabilities, such as attributing the forecasts to either local momenta or global periodicity and characterizing certain core periodic properties, e.g., amplitudes and frequencies. Extensive experiments on both synthetic data and real-world data demonstrate the effectiveness of DEPTS on handling PTS. In most cases, DEPTS achieves significant improvements over the best baseline. Specifically, the error reduction can even reach up to 20% for a few cases. All codes for this paper are publicly available.

****

Automatic Concept Extraction for Concept Bottleneck-based Video Classification

Jeya Vikranth Jeyakumar,Luke Dickens,Yu-Hsi Cheng,Joseph Noor,Luis Antonio Garcia,Diego Ramirez Echavarria,Alessandra Russo,Lance M. Kaplan,Mani Srivastava

Recent efforts in interpretable deep learning models have shown that concept-based explanation methods achieve competitive accuracy with standard end-to-end models and enable reasoning and intervention about extracted high-level visual concepts from images, e.g., identifying the wing color and beak length for bird-species classification. However, these concept bottleneck models rely on a domain expert providing a necessary and sufficient set of concepts--which is intractable for complex tasks such as video classification. For complex tasks, the labels and the relationship between visual elements span many frames, e.g., identifying a bird flying or catching prey--necessitating concepts with various levels of abstraction. To this end, we present CoDEx, an automatic Concept Discovery and Extraction module that rigorously composes a necessary and sufficient set of concept abstractions for concept-based video classification. CoDEx identifies a rich set of complex concept abstractions from natural language explanations of videos--obviating the need to predefine the amorphous set of concepts. To demonstrate our method’s viability, we construct two new public datasets that combine existing complex video classification datasets with short, crowd-sourced natural language explanations for their labels. Our method elicits inherent complex concept abstractions in natural language to generalize concept-bottleneck methods to complex tasks.

****

On Robust Prefix-Tuning for Text Classification

Zonghan Yang,Yang Liu

Recently, prefix-tuning has gained increasing attention as a parameter-efficient finetuning method for large-scale pretrained language models. The method keeps the pretrained models fixed and only updates the prefix token parameters for each downstream task. Despite being lightweight and modular, prefix-tuning still lacks robustness to textual adversarial attacks. However, most currently developed defense techniques necessitate auxiliary model update and storage, which inevitably hamper the modularity and low storage of prefix-tuning. In this work, we propose a robust prefix-tuning framework that preserves the efficiency and modularity of prefix-tuning. The core idea of our framework is leveraging the layerwise activations of the language model by correctly-classified training data as the standard for additional prefix finetuning. During the test phase, an extra batch-level prefix is tuned for each batch and added to the original prefix for robustness enhancement. Extensive experiments on three text classification benchmarks show that our framework substantially improves robustness over several strong baselines against five textual attacks of different types while maintaining comparable accuracy on clean texts. We also interpret our robust prefix-tuning framework from the optimal control perspective and pose several directions for future research.

****

MLP-based architecture with variable length input for automatic speech recognition

Jin Sakuma,Tatsuya Komatsu,Robin Scheibler

We propose multi-layer perceptron (MLP)-based architectures suitable for variable length input.

Recently, several such architectures that do not rely on self-attention have been proposed for image classification.
They achieve performance competitive with that of transformer-based architectures, albeit with a simpler structure and low computational cost.
They split an image into patches and mix information by applying MLPs within and across patches alternately.
Due to the use of MLPs, one such model can only be used for inputs of a fixed, pre-defined size.
However, many types of data are naturally variable in length, for example acoustic signals.
We propose three approaches to extend MLP-based architectures for use with sequences of arbitrary length.
In all of them, we start by splitting the signal into contiguous tokens of fixed size (equivalent to patches in images).
Naturally, the number of tokens is variable.
The two first approaches use a gating mechanism that mixes local information across tokens in a shift-invariant and length-agnostic way.
One uses a depthwise convolution to derive the gate values, while the other rely on shifting tokens.
The final approach explores non-gated mixing using a circular convolution applied in the Fourier domain.
We evaluate the proposed architectures on an automatic speech recognition task with the Librispeech and Tedlium2 corpora. Compared to Transformer, our proposed architecture reduces the WER by \SI{1.9 / 3.4}{\percent} on Librispeech test-clean/test-other set, and 1.8 / 1.6 % on Tedlium2 dev/test set, using only 75.3 % of the parameters.
****
Natural Attribute-based Shift Detection
Jeonghoon Park,Jimin Hong,Radhika Dua,Daehoon Gwak,Jaegul Choo,Sharon Li,Edward Choi
Despite the impressive performance of deep networks in vision, language, and healthcare, unpredictable behaviors on samples from the distribution different than the training distribution cause severe problems in deployment. For better reliability of neural-network-based classifiers, we define a new task, natural attribute-based shift (NAS) detection, to detect the samples shifted from the training distribution by some natural attribute such as age of subjects or brightness of images. Using the natural attributes present in existing datasets, we introduce benchmark datasets in vision, language, and medical for NAS detection. Further, we conduct an extensive evaluation of prior representative out-of-distribution (OOD) detection methods on NAS datasets and observe an inconsistency in their performance. To understand this, we provide an analysis on the relationship between the location of NAS samples in the feature space and the performance of distance- and confidence-based OOD detection methods. Based on the analysis, we split NAS samples into three categories and further suggest a simple modification to the training objective to obtain an improved OOD detection method that is capable of detecting samples from all NAS categories.
****
Learning Graphon Mean Field Games and Approximate Nash Equilibria
Kai Cui,Heinz Koeppl
Recent advances at the intersection of dense large graph limits and mean field games have begun to enable the scalable analysis of a broad class of dynamical sequential games with large numbers of agents. So far, results have been largely limited to graphon mean field systems with continuous-time diffusive or jump dynamics, typically without control and with little focus on computational methods. We propose a novel discrete-time formulation for graphon mean field games as the limit of non-linear dense graph Markov games with weak interaction. On the theoretical side, we give extensive and rigorous existence and approximation properties of the graphon mean field solution in sufficiently large systems. On the practical side we provide general learning schemes for graphon mean field equilibria by either introducing agent equivalence classes or reformulating the graphon m

ean field system as a classical mean field system. By repeatedly finding a regularized optimal control solution and its generated mean field, we successfully obtain plausible approximate Nash equilibria in otherwise infeasible large dense graph games with many agents. Empirically, we are able to demonstrate on a number of examples that the finite-agent behavior comes increasingly close to the mean field behavior for our computed equilibria as the graph or system size grows, verifying our theory. More generally, we successfully apply policy gradient reinforcement learning in conjunction with sequential Monte Carlo methods.

****

Towards Physical, Imperceptible Adversarial Attacks via Adversarial Programs

Itai Mesery,Dana Drachsler Cohen

Adversarial examples were originally defined as imperceptible perturbations which cause a deep neural network to misclassify. However, the majority of imperceptible perturbation attacks require to perturb a large number of pixels across the image and are thus hard to execute in the physical world. Existing physical attacks rely on physical objects, such as patches/stickers or 3D-printed objects. Producing adversarial patches is arguably easier than 3D-printing but these attacks incur highly visible perturbations. This raises the question: is it possible to generate adversarial examples with imperceptible patches? In this work, we consider adversarial multi-patch attacks, where the goal is to compute a targeted attack consisting of up to K patches with minimal L2 distortion. Each patch is associated with dimensions, position, and perturbation parameters. We leverage ideas from program synthesis and numerical optimization to search in this large, discrete space and obtain attacks that are competitive with the C&W attack but have at least 3x and up to 10x fewer perturbed pixels. We evaluate our approach on MNIST, Fashion-MNIST, CIFAR-10, and ImageNet and obtain success rate of at least 92% and up to 100% with at most ten patches. For MNIST, Fashion-MNIST, and CIFAR-10, the average L2 distortion is greater than the average L2 distortion of the C&W attack by up to 1.2.

****

D^2-GCN: Data-Dependent GCNs for Boosting Both Efficiency and Scalability

Chaojian Li,Xu Ouyang,Yang Zhao,Haoran You,Yonggan Fu,Yuchen Gu,Haonan Liu,Siyuan Miao,Yingyan Lin

Graph Convolutional Networks (GCNs) have gained an increasing attention thanks to their state-of-the-art (SOTA) performance in graph-based learning tasks. However, their sheer number of node features and large adjacency matrix limit their deployment into real-world applications, as they impose the following challenges: (1) prohibitive inference cost, especially for resource-constrained applications and (2) low trainability of deep GCNs. To this end, we aim to develop low-cost GCNs with improved trainability, as inspired by recent findings in deep neural network optimization which show that not all data/(model components) are equally important. Specifically, we propose a Data-Dependent GCN framework dubbed D^2-GCN which integrates data-dependent dynamic skipping at multiple granularities: (1) node-wise skipping to bypass aggregating features of unimportant neighbor nodes and their corresponding combinations; (2) edge-wise skipping to prune the unimportant edge connections of each node; and (3) bit-wise skipping to dynamically adapt the bit-precision of both the node features and weights. Our D^2-GCN is achieved by identifying the importance of node features via a low-cost indicator, and thus is simple and generally applicable to various graph-based learning tasks. Extensive experiments and ablation studies on 6 GCN model and dataset pairs consistently validate that the proposed D^2-GCN can (1) largely squeeze out unnecessary costs from both the aggregation and combination phases (e.g., reduce the inference FLOPs by $\downarrow$1.1$\times$ $\sim$ $\downarrow$37.0$\times$ and shrink the energy cost of GCN inference by $\downarrow$1.6$\times$ $\sim$ $\downarrow$8.4$\times$), while offering a comparable or an even better accuracy (e.g., $\downarrow$ 0.5% $\sim$ $\uparrow$ 5.6%); and (2) help GCNs to go deeper by boosting their trainability (e.g., providing a $\uparrow$ 0.8% $\sim$ $\uparrow$ 5.1% higher accuracy when increasing the model depth from 4 layers to 64 layers) and thus achieving a comparable or even better accuracy of GCNs with more layers over SOTA techniques (e.g., a $\downarrow$0.4% $\sim$ $\uparrow$38.6% hi gher accuracy for models with 64 layers). All the codes and pretrained models will be released upon acceptance.

****

AF$_2$: Adaptive Focus Framework for Aerial Imagery Segmentation

Lin Huang,Qiyuan Dong,Jia Zhang,Lijun Wu,Jiang Bian,Tie-Yan Liu

As a specific semantic segmentation task, aerial imagery segmentation has been widely employed in high spatial resolution (HSR) remote sensing images understanding. Besides common issues (e.g. large scale variation) faced by general semantic segmentation tasks, aerial imagery segmentation has some unique challenges, the most critical one among which lies in foreground-background imbalance. There have been some recent efforts that attempt to address this issue by proposing sophisticated neural network architectures, since they can be used to extract informative multi-scale feature representations and increase the discrimination of object boundaries. Nevertheless, many of them merely utilize those multi-scale representations in ad-hoc measures but disregard the fact that the semantic meaning of objects with various sizes could be better identified via receptive fields of diverse ranges. In this paper, we propose Adaptive Focus Framework (AF$_2$), which adopts a hierarchical segmentation procedure and focuses on adaptively utilizing multi-scale representations generated by widely adopted neural network architectures. Particularly, a learnable module, called Adaptive Confidence Mechanism (ACM), is proposed to determine which scale of representation should be used for the segmentation of different objects. Comprehensive experiments show that AF$_2$ has significantly improved the accuracy on three widely used aerial benchmarks, as fast as the mainstream method.

****

BCDR: Betweenness Centrality-based Distance Resampling for Graph Shortest Distance Embedding

Haoyu Wang,Chun Yuan

Along with unprecedented development in network analysis such as biomedical structure prediction and social relationship analysis, Shortest Distance Queries (SDQs) in graphs receive an increasing attention. Approximate algorithms of SDQs with reduced complexity are of vital importance to complex graph applications. Among different approaches, embedding-based distance prediction has made a breakthrough in both efficiency and accuracy, ascribing to the significant performance of Graph Representation Learning (GRL). Embedding-based distance prediction usually leverages truncated random walk followed by Pointwise Mutual Information (PMI)-based optimization to embed local structural features into a dense vector on each node and integrates with a subsequent predictor for global extraction of nodes' mutual shortest distance. It has several shortcomings. Random walk as an unstrained node sequence possesses a limited distance exploration, failing to take into account remote nodes under graph's shortest distance metric, while the PMI-based maximum likelihood optimization of node embeddings reflects excessively versatile local similarity, which incurs an adverse impact on the preservation of the exact shortest distance relation during the mapping from the original graph space to the embedded vector space.

■■

To address these shortcomings, we propose in this paper a novel graph shortest distance embedding method called Betweenness Centrality-based Distance Resampling (BCDR). First, we prove in a statistical perspective that Betweenness Centrality(BC)-based random walk can occupy a wider distance range measured by the intrinsic metric in the graph domain due to its awareness of the path structure. Second, we perform Distance Resampling (DR) from original walk paths before maximum likelihood optimization instead of the PMI-based optimization and prove that this strategy preserves distance relation with respect to any calibrated node via steering optimization objective to reconstruct a global distance matrix. Our proposed method possesses a strong theoretical background and shows much better performance than existing methods when evaluated on a broad class of real-world graph datasets with large diameters in SDQ problems. It should also outperform existing methods in other graph structure-related applications.

****

Measuring CLEVRness: Black-box Testing of Visual Reasoning Models

Spyridon Mouselinos,Henryk Michalewski,Mateusz Malinowski

How can we measure the reasoning capabilities of intelligence systems? Visual question answering provides a convenient framework for testing the model's abilities by interrogating the model through questions about the scene. However, despite scores of various visual QA datasets and architectures, which sometimes yield even a super-human performance, the question of whether those architectures can actually reason remains open to debate.

To answer this, we extend the visual question answering framework and propose the following behavioral test in the form of a two-player game. We consider black-box neural models of CLEVR. These models are trained on a diagnostic dataset benchmarking reasoning. Next, we train an adversarial player that re-configures the scene to fool the CLEVR model. We show that CLEVR models, which otherwise could perform at a ``human-level'', can easily be fooled by our agent. Our results put in doubt whether data-driven approaches can do reasoning without exploiting the numerous biases that are often present in those datasets. Finally, we also propose a controlled experiment measuring the efficiency of such models to learn and perform reasoning.

****

Tractable Dendritic RNNs for Identifying Unknown Nonlinear Dynamical Systems

Manuel Brenner,Leonard Bereska,Jonas Magdy Mikhaeil,Florian Hess,Zahra Monfared,Po-Chen Kuo,Daniel Durstewitz

In many scientific disciplines, we are interested in inferring the nonlinear dynamical system underlying a set of observed time series, a challenging task in the face of chaotic behavior and noise. Previous deep learning approaches toward this goal often suffered from a lack of interpretability and tractability. In particular, the high-dimensional latent spaces often required for a faithful embedding, even when the underlying dynamics lives on a lower-dimensional manifold, can hamper theoretical analysis. Motivated by the emerging principles of dendritic computation, we augment a dynamically interpretable and mathematically tractable piecewise-linear (PL) recurrent neural network (RNN) by a linear spline basis expansion. We show that this approach retains all the theoretically appealing properties of the simple PLRNN, yet boosts its capacity for approximating arbitrary nonlinear dynamical systems in comparatively low dimensions. We introduce two frameworks for training the system, one based on fast and scalable variational inference, and another combining BPTT with teacher forcing. We show that the dendritically expanded PLRNN achieves better reconstructions with fewer parameters and dimensions on various dynamical systems benchmarks and compares favorably to other methods, while retaining a tractable and interpretable structure.

****

Sparse Hierarchical Table Ensemble

Guy Farjon,Aharon Bar HIllel

Deep learning for tabular data is drawing increasing attention, with recent work attempting to boost the accuracy of neuron-based networks. However, when computational capacity is low as in Internet of Things (IoT), drone, or Natural User Interface (NUI) applications, such deep learning methods are deserted. We offer to enable deep learning capabilities using ferns (oblivious decision trees) instead of neurons, by constructing a Sparse Hierarchical Table Ensemble (S-HTE). S-HTE inference is dense at the beginning of the training process and becomes gradually sparse using an annealing mechanism, leading to an efficient final predictor. Unlike previous work with ferns, S-HTE learns useful internal representations, and it earns from increasing depth. Using a standard classification and regression benchmark, we show its accuracy is comparable to alternatives while having an order of magnitude lower computational complexity. Our PyTorch implementation is available at https://anonymous.4open.science/r/HTE_CTE-60EB/

****

Combining Differential Privacy and Byzantine Resilience in Distributed SGD

Rachid Guerraoui,Nirupam Gupta,Rafael Pinot,Sébastien Rouault,John Stephan

Privacy and Byzantine resilience (BR) are two crucial requirements of modern-day distributed machine learning. The two concepts have been extensively studied individually but the question of how to combine them effectively remains unanswered. This paper contributes to addressing this question by studying the extent to which the distributed SGD algorithm, in the standard parameter-server architecture, can learn an accurate model despite (a) a fraction of the workers being malicious (Byzantine), and (b) the other fraction, whilst being honest, providing noisy information to the server to ensure differential privacy (DP). We first observe that the integration of standard practices in DP and BR is not straightforward. In fact, we show that many existing results on the convergence of distributed SGD under Byzantine faults, especially those relying on (α,f)-Byzantine resilience, are rendered invalid when honest workers enforce DP. To circumvent this shortcoming, we revisit the theory of (α,f)-BR to obtain an approximate convergence guarantee. Our analysis provides key insights on how to improve this guarantee through hyperparameter optimization. Essentially, our theoretical and empirical results show that (1) an imprudent combination of standard approaches to DP and BR might be fruitless, but (2) by carefully re-tuning the learning algorithm, we can obtain reasonable learning accuracy while simultaneously guaranteeing DP and BR.

****

When high-performing models behave poorly in practice: periodic sampling can help

Stanislas Chambon,Julien GUILLAUMIN,Luis Montero,Yaroslav Nikulin,Paul Wambergue,Pierre Fillard

Training a deep neural network (DNN) for breast cancer detection from medical images suffers from the (hopefully) low prevalence of the pathology.

For a sensible amount of positive cases, images must be collected from numerous places resulting in large heterogeneous datasets with different acquisition devices, populations, cancer incidences.

Without precaution, this heterogeneity may result in a DNN biased by latent variables a priori independent of the pathology.

This may be dramatic if this DNN is used inside a software to help radiologists to detect cancers.

This work mitigates this issue by acting on how mini-batches for Stochastic Gradient Descent (SGD) algorithms are constructed.

The dataset is divided into homogeneous subsets sharing some attributes (\textit{e.g.} acquisition device, source) called Data Segments (DSs).

Batches are built by sampling each DS periodically with a frequency proportional to the rarest label in the DS and by simultaneously preserving an overall balance between positive and negative labels within the batch.

Periodic sampling is compared to balanced sampling (equal amount of labels within a batch, independently of DS) and to balanced sampling within DS (equal amount of labels within a batch and each DS).

We show, on breast cancer prediction from mammography images of various devices and origins, that periodic sampling leads to better generalization than other sampling strategies.

****

Align-RUDDER: Learning From Few Demonstrations by Reward Redistribution

Vihang Prakash Patil,Markus Hofmarcher,Marius-Constantin Dinu,Matthias Dorfer,Patrick M Blies,Johannes Brandstetter,Jose Arjona-Medina,Sepp Hochreiter

Reinforcement Learning algorithms require a large number of samples to solve complex tasks with sparse and delayed rewards. Complex tasks are often hierarchically composed of sub-tasks. Solving a sub-task increases the return expectation and leads to a step in the Q-function. RUDDER identifies these steps and then redistributes reward to them, thus immediately giving reward if sub-tasks are solved. Since the delay of rewards is reduced, learning is considerably sped up. However, for complex tasks, current exploration strategies struggle with discovering episodes with high rewards. Therefore, we assume that episodes with high rewards are given as demonstrations and do not have to be discovered by exploration. Unfortunately, the number of demonstrations is typically small and RUDDER's LSTM as a deep learning model does not learn well on these few training samples. Hen ce, we introduce Align-RUDDER, which is RUDDER with two major modifications. First, Align-RUDDER assumes that episodes with high rewards are given as demonstrations, replacing RUDDER’s safe exploration and lessons replay buffer. Second, we substitute RUDDER’s LSTM model by a profile model that is obtained from multiple sequence alignment of demonstrations. Profile models can be constructed from as few as two demonstrations. Align-RUDDER uses reward redistribution to speed up learning by reducing the delay of rewards. Align-RUDDER outperforms competitors on complex artificial tasks with delayed rewards and few demonstrations. On the MineCraft ObtainDiamond task, Align-RUDDER is able to mine a diamond, though not frequently.

****

p-Laplacian Based Graph Neural Networks

Guoji Fu,Peilin Zhao,Yatao Bian

Graph neural networks (GNNs) have demonstrated superior performance for semi-supervised node classification on graphs, as a result of their ability to exploit node features and topological information simultaneously. However, most GNNs implicitly assume that the labels of nodes and their neighbors in a graph are the same or consistent, which does not hold in heterophilic graphs, where the labels of linked nodes are likely to differ. Hence, when the topology is non-informative for label prediction, ordinary GNNs may work significantly worse than simply applying multi-layer perceptrons (MLPs) on each node. To tackle the above problem, we propose a new p-Laplacian based GNN model, termed as ^{p}GNN, whose message passing mechanism is derived from a discrete regularization framework and could be theoretically explained as an approximation of a polynomial graph filter defined on the spectral domain of p-Laplacians. The spectral analysis shows that the new message passing mechanism works simultaneously as low-pass and high-pass filters, thus making ^{p}GNNs effective on both homophilic and heterophilic graphs. Empirical studies on real-world and synthetic datasets validate our findings and demonstrate that ^{p}GNNs significantly outperform several state-of-the-art GNN architectures on heterophilic benchmarks while achieving competitive performance on homophilic benchmarks. Moreover, ^{p}GNNs can adaptively learn aggregation weights and are robust to noisy edges.

****

Learning Diverse Options via InfoMax Termination Critic

Yuji Kanagawa,Tomoyuki Kaneko

We consider the problem of autonomously learning reusable temporally extended actions, or options, in reinforcement learning. While options can speed up transfer learning by serving as reusable building blocks, learning reusable options for unknown task distribution remains challenging. Motivated by the recent success of mutual information (MI) based skill learning, we hypothesize that more diverse options are more reusable. To this end, we propose a method for learning termination conditions of options by maximizing MI between options and corresponding state transitions. We derive a scalable approximation of this MI maximization via gradient ascent, yielding the InfoMax Termination Critic (IMTC) algorithm. Our experiments demonstrate that IMTC significantly improves the diversity of learned options without rewards, combined with an intrinsic option learning method. Moreover, we test the reusability of learned options by transferring options into various tasks, confirming that IMTC helps quick adaptation, especially in complex domains where an agent needs to manipulate objects.

****

Understanding and Scheduling Weight Decay

Zeke Xie,Issei Sato,Masashi Sugiyama

Weight decay is a popular and even necessary regularization technique for training deep neural networks that generalize well. Previous work usually interpreted weight decay as a Gaussian prior from the Bayesian perspective. However, weight decay sometimes shows mysterious behaviors beyond the conventional understanding . For example, the optimal weight decay value tends to be zero given long enough training time. Moreover, existing work typically failed to recognize the importance of scheduling weight decay during training. Our work aims at theoretically understanding novel behaviors of weight decay and designing schedulers for weigh

t decay in deep learning. This paper mainly has three contributions. First, we propose a novel theoretical interpretation of weight decay from the perspective of learning dynamics. Second, we propose a novel weight-decay linear scaling rule for large-batch training that proportionally increases weight decay rather than the learning rate as the batch size increases. Third, we provide an effective learning-rate-aware scheduler for weight decay, called the Stable Weight Decay (SWD) method, which, to the best of our knowledge, is the first practical design for weight decay scheduling. In our various experiments, the SWD method often makes improvements over L_{2} Regularization and Decoupled Weight Decay.
****

Evaluation Metrics for Graph Generative Models: Problems, Pitfalls, and Practical Solutions

Leslie O'Bray,Max Horn,Bastian Rieck,Karsten Borgwardt

Graph generative models are a highly active branch of machine learning. Given the steady development of new models of ever-increasing complexity, it is necessary to provide a principled way to evaluate and compare them. In this paper, we enumerate the desirable criteria for such a comparison metric and provide an overview of the status quo of graph generative model comparison in use today, which predominantly relies on the maximum mean discrepancy (MMD). We perform a systematic evaluation of MMD in the context of graph generative model comparison, highlighting some of the challenges and pitfalls researchers inadvertently may encounter. After conducting a thorough analysis of the behaviour of MMD on synthetically-generated perturbed graphs as well as on recently-proposed graph generative models, we are able to provide a suitable procedure to mitigate these challenges and pitfalls. We aggregate our findings into a list of practical recommendations for researchers to use when evaluating graph generative models.
****

Exploiting Class Activation Value for Partial-Label Learning

Fei Zhang,Lei Feng,Bo Han,Tongliang Liu,Gang Niu,Tao Qin,Masashi Sugiyama

Partial-label learning (PLL) solves the multi-class classification problem, where each training instance is assigned a set of candidate labels that include the true label. Recent advances showed that PLL can be compatible with deep neural networks, which achieved state-of-the-art performance. However, most of the existing deep PLL methods focus on designing proper training objectives under various assumptions on the collected data, which may limit their performance when the collected data cannot satisfy the adopted assumptions. In this paper, we propose to exploit the learned intrinsic representation of the model to identify the true label in the training process, which does not rely on any assumptions on the collected data. We make two key contributions. As the first contribution, we empirically show that the class activation map (CAM), a simple technique for discriminating the learning patterns of each class in images, could surprisingly be utilized to make accurate predictions on selecting the true label from candidate labels. Unfortunately, as CAM is confined to image inputs with convolutional neural networks, we are yet unable to directly leverage CAM to address the PLL problem with general inputs and models. Thus, as the second contribution, we propose the class activation value (CAV), which owns similar properties of CAM, while CAV is versatile in various types of inputs and models. Building upon CAV, we propose a novel method named CAV Learning (CAVL) that selects the true label by the class with the maximum CAV for model training. Extensive experiments on various datasets demonstrate that our proposed CAVL method achieves state-of-the-art performance.
****

Natural Posterior Network: Deep Bayesian Predictive Uncertainty for Exponential Family Distributions

Bertrand Charpentier,Oliver Borchert,Daniel Zügner,Simon Geisler,Stephan Günnemann

Uncertainty awareness is crucial to develop reliable machine learning models. In this work, we propose the Natural Posterior Network (NatPN) for fast and high-quality uncertainty estimation for any task where the target distribution belongs to the exponential family. Thus, NatPN finds application for both classificatio n and general regression settings. Unlike many previous approaches, NatPN does not require out-of-distribution (OOD) data at training time. Instead, it leverages Normalizing Flows to fit a single density on a learned low-dimensional and task-dependent latent space. For any input sample, NatPN uses the predicted likelihood to perform a Bayesian update over the target distribution. Theoretically, NatPN assigns high uncertainty far away from training data. Empirically, our extensive experiments on calibration and OOD detection show that NatPN delivers highly competitive performance for classification, regression and count prediction tasks.

****

What to expect of hardware metric predictors in NAS

Kevin Alexander Laube,Maximus Mutschler,Andreas Zell

Modern Neural Architecture Search (NAS) focuses on finding the best performing architectures in hardware-aware settings; e.g., those with an optimal tradeoff of accuracy and latency. Due to many advantages of prediction models over live measurements, the search process is often guided by estimates of how well each considered network architecture performs on the desired metrics. Typical prediction models range from operation-wise lookup tables over gradient-boosted trees and neural networks, with little known information on how they compare. We evaluate 18 different performance predictors on ten combinations of metrics, devices, network types, and training tasks, and find that MLP models are the most promising. We then simulate and evaluate how the guidance of such prediction models affects the subsequent architecture selection. Due to inaccurate predictions, the selected architectures are generally suboptimal, which we quantify as an expected reduction in accuracy and hypervolume. We show that simply verifying the predictions of just the selected architectures can lead to substantially improved results. Under a time budget, we find it preferable to use a fast and inaccurate prediction model over accurate but slow live measurements.

****

Givens Coordinate Descent Methods for Rotation Matrix Learning in Trainable Embedding Indexes

Yunjiang Jiang,Han Zhang,Yiming Qiu,Yun Xiao,Bo Long,Wen-Yun Yang

Product quantization (PQ) coupled with a space rotation, is widely used in modern approximate nearest neighbor (ANN) search systems to significantly compress the disk storage for embeddings and speed up the inner product computation. Existing rotation learning methods, however, minimize quantization distortion for fixed embeddings, which are not applicable to an end-to-end training scenario where embeddings are updated constantly. In this paper, based on geometric intuitions from Lie group theory, in particular the special orthogonal groupSO(n), we propose a family of block Givens coordinate descent algorithms to learn rotation matrix that are provably convergent on any convex objectives. Compared to the state-of-the-art SVD method, the Givens algorithms are much more parallelizable, reducing runtime by orders of magnitude on modern GPUs, and converge more stably according to experimental studies. They further improve upon vanilla product quantization significantly in an end-to-end training scenario.

****

An Attempt to Model Human Trust with Reinforcement Learning

Vincent Frey,Simon Bécot

Existing works to compute trust as a numerical value mainly rely on ranking, rating or assessments of agents by other agents. However, the concept of trust is manifold, and should not be limited to reputation. Recent research in neuroscience converges with Berg's hypothesis in economics that trust is an encoded function in the human brain. Based on this new assumption, we propose an approach where a trust level is learned by an overlay of any model-free off-policy reinforcement learning algorithm. The main issues were i) to use recent findings on dopaminergic system and reward circuit to simulate trust, ii) to assess our model with reliable and unbiased real life models. In this work, we address these problems by extending Q-Learning to trust evaluation, and comparing our results to a social science case study. Our main contributions are threefold. (1) We model the trust-decision making process with a reinforcement learning algorithm. (2) We prop ose a dynamic reinforcement of the trust reward inspired by recent findings of neuroscience. (3) We propose a method to explore and exploit the trust space. The experiments reveal that it is possible to find a set of hyperparameters of our algorithm to reproduce recent findings on overconfidence effect in social psychology research.

****

Beyond Quantization: Power aware neural networks

Nurit Spingarn,Elad Hoffer,Ron Banner,Hilla Ben Yaacov,Tomer Michaeli

Power consumption is a major obstacle in the deployment of deep neural networks (DNNs) on end devices. Existing approaches for reducing power consumption rely on quite general principles, including avoidance of multiplication operations and aggressive quantization of weights and activations. However, these methods do not take into account the precise power consumed by each module in the network, and are therefore far from optimal. In this paper we develop accurate power consumption models for all arithmetic operations in the DNN, under various working conditions. Surprisingly, we reveal several important factors that have been overlooked to date. Based on our analysis, we present PANN (power-aware neural network), a simple approach for approximating any full-precision network by a low-power fixed-precision variant. Our method can be applied to a pre-trained network, and can also be used during training to achieve improved performance. In contrast to previous approaches, our method incurs only a minor degradation in accuracy w.r.t. the full-precision version of the network, even when working at the power-budget of a 2-bit quantized variant. In addition, our scheme enables to seamlessly traverse the power-accuracy tradeoff at deployment time, which is a major advantage over existing quantization methods that are constrained to specific bit widths.

****

Can Stochastic Gradient Langevin Dynamics Provide Differential Privacy for Deep Learning?

Guy Heller,Ethan Fetaya

Bayesian learning via Stochastic Gradient Langevin Dynamics (SGLD) has been suggested for differentially private learning. While previous research provides differential privacy bounds for SGLD when close to convergence or at the initial steps of the algorithm, the question of what differential privacy guarantees can be made in between remains unanswered. This interim region is essential, especially for Bayesian neural networks, as it is hard to guarantee convergence to the posterior. This paper will show that using SGLD might result in unbounded privacy loss for this interim region, even when sampling from the posterior is as differentially private as desired.

****

SPIDE: A Purely Spike-based Method for Training Feedback Spiking Neural Networks

Mingqing Xiao,Qingyan Meng,Zongpeng Zhang,Yisen Wang,Zhouchen Lin

Spiking neural networks (SNNs) with event-based computation are promising brain-inspired models for energy-efficient applications on neuromorphic hardware. However, most supervised SNN training methods require complex computation or impractical neuron models, which hinders them from spike-based energy-efficient training. Among them, the recently proposed method, implicit differentiation on the equilibrium state (IDE), for training feedback SNNs is a promising way that is possible for generalization to locally spike-based learning with flexible network structures. In this paper, we study spike-based implicit differentiation on the equilibrium state (SPIDE) that extends the IDE method for supervised local learning with spikes, which could be possible for energy-efficient training on neuromorphic hardware. Specifically, we first introduce ternary spiking neuron couples that can realize ternary outputs with the common neuron model, and we prove that implicit differentiation can be solved by spikes based on this design. With this approach, the whole training procedure can be made as event-driven spike computation and weights are updated locally with two-stage average firing rates. Then to reduce the approximation error of spikes due to the finite simulation time steps, we propose to modify the resting membrane potential. Based on it, the average firing rate, when viewed as a stochastic estimator, achieves an unbiased esti mation of iterative solution for implicit differentiation and the variance of this estimator is reduced. With these key components, we can train SNNs with either feedback or feedforward structures in a small number of time steps. Further, the firing sparsity during training demonstrates the great potential for energy efficiency. Meanwhile, even with these constraints, our trained models could still achieve competitive results on MNIST, CIFAR-10 and CIFAR-100. Our proposed method demonstrates the great potential for energy-efficient training of SNNs on neuromorphic hardware.

****

Few-shot Learning with Big Prototypes

Ning Ding,Yulin Chen,Xiaobin Wang,Hai-Tao Zheng,Zhiyuan Liu,Pengjun Xie

Using dense vectors, i.e., prototypes, to represent abstract information of classes has become a common approach in low-data machine learning scenarios. Typically, prototypes are mean output embeddings over the instances for each class. In this case, prototypes have the same dimension of example embeddings, and such tensors could be regarded as ``points'' in the feature space from the geometrical perspective. But these points may lack the expressivity of the whole class-level information due to the biased sampling.

In this paper, we propose to use tensor fields (``areas'') to model prototypes to enhance the expressivity of class-level information. Specifically, we present \textit{big prototypes}, where prototypes are represented by hyperspheres with dynamic sizes. A big prototype could be effectively modeled by two sets of learnable parameters, one is the center of the hypersphere, which is an embedding with the same dimension of training examples. The other is the radius of the sphere, which is a constant. Compared with atactic manifolds with complex boundaries, representing hypersphere with parameters is immensely easier. Moreover, it is convenient to perform metric-based classification with big prototypes in few-shot learning, where we only need to calculate the distance from a data point to the surface of the hypersphere.

Extensive experiments on few-shot learning tasks across NLP and CV demonstrate the effectiveness of big prototypes.

****

cosFormer: Rethinking Softmax In Attention

Zhen Qin,Weixuan Sun,Hui Deng,Dongxu Li,Yunshen Wei,Baohong Lv,Junjie Yan,Lingpeng Kong,Yiran Zhong

Transformer has shown great successes in natural language processing, computer vision, and audio processing. As one of its core components, the softmax attention helps to capture long-range dependencies yet prohibits its scale-up due to the quadratic space and time complexity to the sequence length. Kernel methods are often adopted to reduce the complexity by approximating the softmax operator. Nevertheless, due to the approximation errors, their performances vary in different tasks/corpus and suffer crucial performance drops when compared with the vanilla softmax attention. In this paper, we propose a linear transformer called cosFormer that can achieve comparable or better accuracy to the vanilla transformer in both casual and cross attentions. cosFormer is based on two key properties of softmax attention: i). non-negativeness of the attention matrix; ii). a non-linear re-weighting scheme that can concentrate the distribution of the attention matrix. As its linear substitute, cosFormer fulfills these properties with a linear operator and a cosine-based distance re-weighting mechanism. Extensive experiments on language modeling and text understanding tasks demonstrate the effectiveness of our method. We further examine our method on long sequences and achieve state-of-the-art performance on the Long-Range Arena benchmark. The source code is available at https://github.com/OpenNLPLab/cosFormer.

****

Safe Opponent-Exploitation Subgame Refinement

Mingyang Liu,Chengjie Wu,Qihan Liu,Yansen Jing,Jun Yang,Pingzhong Tang,Chongjie Zhang

Search algorithms have been playing a vital role in the success of superhuman AI in both perfect information and imperfect information games. Specifically, sear ch algorithms can generate a refinement of Nash equilibrium (NE) approximation in games such as Texas hold'em with theoretical guarantees. However, when confronted with opponents of limited rationality, an NE strategy tends to be overly conservative, because it prefers to achieve its low exploitability rather than actively exploiting the weakness of opponents. In this paper, we investigate the dilemma of safety and opponent exploitation. We present a new real-time search framework that smoothly interpolates between the two extremes of strategy search, hence unifying safe search and opponent exploitation. We provide our new strategy with a theoretically upper-bounded exploitability and lower-bounded reward against an opponent. Our method can exploit the weakness of its opponent without significantly sacrificing its exploitability. Empirical results show that our method significantly outperforms NE baselines when opponents play non-NE strategies and keeps low exploitability at the same time.

****

FALCON: Fast Visual Concept Learning by Integrating Images, Linguistic descriptions, and Conceptual Relations
Lingjie Mei,Jiayuan Mao,Ziqi Wang,Chuang Gan,Joshua B. Tenenbaum
We present a meta-learning framework for learning new visual concepts quickly, from just one or a few examples, guided by multiple naturally occurring data streams: simultaneously looking at images, reading sentences that describe the objects in the scene, and interpreting supplemental sentences that relate the novel concept with other concepts. The learned concepts support downstream applications, such as answering questions by reasoning about unseen images. Our model, namely FALCON, represents individual visual concepts, such as colors and shapes, as axis-aligned boxes in a high-dimensional space (the ``box embedding space''). Given an input image and its paired sentence, our model first resolves the referential expression in the sentence and associates the novel concept with particular objects in the scene. Next, our model interprets supplemental sentences to relate the novel concept with other known concepts, such as ``X has property Y'' or ``X is a kind of Y''. Finally, it infers an optimal box embedding for the novel concept that jointly 1) maximizes the likelihood of the observed instances in the image, and 2) satisfies the relationships between the novel concepts and the known ones. We demonstrate the effectiveness of our model on both synthetic and real-world datasets.

****

How to train RNNs on chaotic data?
Zahra Monfared,Jonas Magdy Mikhaeil,Daniel Durstewitz
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. 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 offer an effective yet simple training technique for chaotic data and guidance on how to choose relevant hyperparameters according to the Lyapunov spectrum.

****

Bayesian Learning with Information Gain Provably Bounds Risk for a Robust Adversarial Defense
Bao Gia Doan,Ehsan M Abbasnejad,Damith C. Ranasinghe

In this paper, we present a novel method to learn a Bayesian neural network robust against adversarial attacks. Previous algorithms have shown an adversarially trained Bayesian Neural Network (BNN) provides improved robustness against attacks. However, the learning approach for approximating the multi-modal Bayesian posterior leads to mode collapse with consequential sub-par robustness and under performance of an adversarially trained BNN. Instead, we propose approximating the multi-modal posterior of a BNN to prevent mode collapse and encourage diversity over learned posterior distributions of models to develop a novel adversarial training method for BNNs. Importantly, we conceptualize and formulate information gain (IG) in the adversarial Bayesian learning context and prove, training a BNN with IG bounds the difference between the conventional empirical risk with the risk obtained from adversarial training---our intuition is that information gain from benign and adversarial examples should be the same for a robust BNN. Extensive experimental results demonstrate our proposed algorithm to achieve state-of-the-art performance under strong adversarial attacks.

****

Understanding over-squashing and bottlenecks on graphs via curvature

Jake Topping,Francesco Di Giovanni,Benjamin Paul Chamberlain,Xiaowen Dong,Michael M. Bronstein

Most graph neural networks (GNNs) use the message passing paradigm, in which node features are propagated on the input graph. Recent works pointed to the distortion of information flowing from distant nodes as a factor limiting the efficiency of message passing for tasks relying on long-distance interactions. This phenomenon, referred to as 'over-squashing', has been heuristically attributed to graph bottlenecks where the number of k-hop neighbors grows rapidly with k. We provide a precise description of the over-squashing phenomenon in GNNs and analyze how it arises from bottlenecks in the graph. For this purpose, we introduce a new edge-based combinatorial curvature and prove that negatively curved edges are responsible for the over-squashing issue. We also propose and experimentally test a curvature-based graph rewiring method to alleviate the over-squashing.

****

Meta Attention For Off-Policy Actor-Critic

Jiateng Huang,Wanrong Huang,Long Lan,Dan Wu

Off-Policy Actor-Critic methods can effectively exploit past experiences and thus they have achieved great success in various reinforcement learning tasks. In many image-based and multi-source tasks, attention mechanism has been employed in Actor-Critic methods to improve their sampling efficiency. In this paper, we propose a meta attention method for state-based reinforcement learning tasks, which combines attention mechanism and meta-learning based on the Off-Policy Actor-Critic framework. Unlike previous attention-based work, our meta attention method introduces attention in the actor and the critic of the typical Actor-Critic framework rather than in multiple pixels of an image or multiple information sources. In contrast to existing meta-learning methods, the proposed meta-attention approach is able to function in both the gradient-based training phase and the agent's decision-making process. The experimental results demonstrate the superiority of our meta-attention method in various continuous control tasks, which are based on the Off-Policy Actor-Critic methods including DDPG, TD3, and SAC.

****

HyAR: Addressing Discrete-Continuous Action Reinforcement Learning via Hybrid Action Representation

Boyan Li,Hongyao Tang,YAN ZHENG,Jianye HAO,Pengyi Li,Zhen Wang,Zhaopeng Meng,LI Wang

Discrete-continuous hybrid action space is a natural setting in many practical problems, such as robot control and game AI. However, most previous Reinforcement Learning (RL) works only demonstrate the success in controlling with either discrete or continuous action space, while seldom take into account the hybrid action space. One naive way to address hybrid action RL is to convert the hybrid action space into a unified homogeneous action space by discretization or continualization, so that conventional RL algorithms can be applied. However, this ignores the underlying structure of hybrid action space and also induces the scalabili ty issue and additional approximation difficulties, thus leading to degenerated results. In this paper, we propose Hybrid Action Representation (HyAR) to learn a compact and decodable latent representation space for the original hybrid action space. HyAR constructs the latent space and embeds the dependence between discrete action and continuous parameter via an embedding table and conditional Variantional Auto-Encoder (VAE). To further improve the effectiveness, the action representation is trained to be semantically smooth through unsupervised environmental dynamics prediction. Finally, the agent then learns its policy with conventional DRL algorithms in the learned representation space and interacts with the environment by decoding the hybrid action embeddings to the original action space. We evaluate HyAR in a variety of environments with discrete-continuous action space. The results demonstrate the superiority of HyAR when compared with previous baselines, especially for high-dimensional action spaces.

****

Character Generation through Self-Supervised Vectorization

Gokcen Gokceoglu,Emre Akbas

The prevalent approach in self-supervised image generation is to operate on pixel level representations. While this approach can produce high quality images, it cannot benefit from the simplicity and innate quality of vectorization. Here we present a drawing agent that operates on stroke-level representation of images. At each time step, the agent first assesses the current canvas and decides whether to stop or keep drawing. When a `draw’ decision is made, the agent outputs a program indicating the stroke to be drawn. As a result, it produces a final raster image by drawing the strokes on a canvas, using a minimal number of strokes and dynamically deciding when to stop. We train our agent through reinforcement learning on MNIST and Omniglot datasets for unconditional generation and parsing (reconstruction) tasks. We utilize our parsing agent for exemplar generation and type conditioned concept generation in Omniglot challenge without any further training. We present successful results on all three generation tasks and the parsing task. Crucially, we do not need any stroke-level or vector supervision; we only use raster images for training. Code will be made available upon acceptance.

****

Towards simple time-to-event modeling: optimizing neural networks via rank regression

Hyunjun Lee,Junhyun Lee,Taehwa Choi,Jaewoo Kang,Sangbum Choi

Time-to-event analysis, also known as survival analysis, aims to predict the first occurred event time, conditional on a set of features.

However, the presence of censorship brings much complexity in learning algorithms due to data incompleteness.

Hazard-based models (e.g. Cox's proportional hazards) and accelerated failure time (AFT) models are two popular tools in time-to-event modeling, requiring the proportional hazards and linearity assumptions, respectively.

In addition, AFT models require pre-specified parametric distributional assumptions in most cases.

To alleviate such strict assumptions and improve predictive performance, there have been many deep learning approaches for hazard-based models in recent years.

However, compared to hazard-based methods, AFT-based representation learning has received limited attention in neural network literature, despite its model simplicity and interpretability.

In this work, we introduce a Deep AFT Rank-regression for Time-to-event prediction model (DART), which is a deep learning-based semiparametric AFT model, and propose a l_1-type rank loss function that is more suitable for optimizing neural networks.

Unlike existing neural network-based AFT models, the proposed model is semiparametric in that any distributional assumption is not imposed for the survival time distribution without requiring further hyperparameters or complicated model architectures.

We verify the usefulness of DART via quantitative analysis upon various benchmar k datasets.

The results show that our method has considerable potential to model high-throughput censored time-to-event data.

****

An Effective GCN-based Hierarchical Multi-label classification for Protein Function Prediction

Kyudam Choi,Yurim Lee,Cheongwon Kim

We propose an effective method to improve Protein Function Prediction (PFP) utilizing hierarchical features of Gene Ontology (GO) terms. Our method consists of a language model for encoding the protein sequence and a Graph Convolutional Network (GCN) for representing Go terms. To reflect the hierarchical structure of GO to GCN, we employ node(GO term)-wise representations containing the whole hierarchical information. Our algorithm shows effectiveness in a large-scale graph by expanding the GO graph compared to previous models. Experimental results show that our method outperformed state-of-the-art PFP approaches.

****

How BPE Affects Memorization in Transformers

Eugene Kharitonov,Marco Baroni,Dieuwke Hupkes

Training data memorization in NLP can both be beneficial (e.g., closed-book QA) and undesirable (personal data extraction). In any case, successful model training requires a non-trivial amount of memorization to store word spellings, various linguistic idiosyncrasies and common knowledge. However, little is known about what affects the memorization behavior of NLP models, as the field tends to focus on the equally important question of generalization.

In this work, we demonstrate that the size of the subword vocabulary learned by Byte-Pair Encoding (BPE) greatly affects both ability and tendency of standard Transformer models to memorize training data, even when we control for the number of learned parameters. We find that with a large subword vocabulary size, Transformer models fit random mappings more easily and are more vulnerable to membership inference attacks. Similarly, given a prompt, Transformer-based language models with large subword vocabularies reproduce the training data more often. We conjecture this effect is caused by reduction in the sequences' length that happens as the BPE vocabulary grows. Our findings can allow a more informed choice of hyper-parameters, that is better tailored for a particular use-case.

****

An evaluation of quality and robustness of smoothed explanations

Ahmad Ajalloeian,Seyed-Mohsen Moosavi-Dezfooli,Michalis Vlachos,Pascal Frossard

Explanation methods play a crucial role in helping to understand the decisions of deep neural networks (DNNs) to develop trust that is critical for the adoption of predictive models. However, explanation methods are easily manipulated through visually imperceptible perturbations that generate misleading explanations. The geometry of the decision surface of the DNNs has been identified as the main cause of this phenomenon and several \emph{smoothing} approaches have been proposed to build more robust explanations.

In this work, we provide a thorough evaluation of the quality and robustness of the explanations derived by smoothing approaches. Their different properties are evaluated with extensive experiments, which reveal the settings where the smoothed explanations are better, and also worse than the explanations derived by the common Gradient method. By making the connection with the literature on adversarial attacks, we further show that such smoothed explanations are robust primarily against additive ℓ_p-norm attacks. However, a combination of additive and non-additive attacks can still manipulate these explanations, which reveals shortcomings in their robustness properties.

****

Transductive Universal Transport for Zero-Shot Action Recognition

Pascal Mettes

This work addresses the problem of recognizing action categories in videos for which no training examples are available. The current state-of-the-art enables such a zero-shot recognition by learning universal mappings from videos to a shared semantic space, either trained on large-scale seen actions or on objects. Whil e effective, universal action and object models are biased to their seen categories. Such biases are further amplified due to biases between seen and unseen categories in the semantic space. The amplified biases result in many unseen action categories simply never being selected during inference, hampering zero-shot progress. We seeks to address this limitation and introduce transductive universal transport for zero-shot action recognition. Our proposal is to re-position unseen action embeddings through transduction, \ie by using the distribution of the unlabelled test set. For universal action models, we first find an optimal mapping from unseen actions to the mapped test videos in the shared hyperspherical space. We then define target embeddings as weighted Fr\'echet means, with the weights given by the transport couplings. Finally, we re-position unseen action embeddings along the geodesic between the original and target, as a form of semantic regularization. For universal object models, we outline a weighted transport variant from unseen action embeddings to object embeddings directly. Empirically, we show that our approach directly boosts universal action and object models, resulting in state-of-the-art performance for zero-shot classification and spatio-temporal localization.

****

Learning Altruistic Behaviours in Reinforcement Learning without External Rewards

Tim Franzmeyer,Mateusz Malinowski,Joao F. Henriques

Can artificial agents learn to assist others in achieving their goals without knowing what those goals are? Generic reinforcement learning agents could be trained to behave altruistically towards others by rewarding them for altruistic behaviour, i.e., rewarding them for benefiting other agents in a given situation. Such an approach assumes that other agents' goals are known so that the altruistic agent can cooperate in achieving those goals. However, explicit knowledge of other agents' goals is often difficult to acquire. In the case of human agents, their goals and preferences may be difficult to express fully; they might be ambiguous or even contradictory. Thus, it is beneficial to develop agents that do not depend on external supervision and learn altruistic behaviour in a task-agnostic manner. We propose to act altruistically towards other agents by giving them more choice and allowing them to achieve their goals better. Some concrete examples include opening a door for others or safeguarding them to pursue their objectives without interference. We formalize this concept and propose an altruistic agent that learns to increase the choices another agent has by preferring to maximize the number of states that the other agent can reach in its future. We evaluate our approach in three different multi-agent environments where another agent's success depends on altruistic behaviour. Finally, we show that our unsupervised agents can perform comparably to agents explicitly trained to work cooperatively, in some cases even outperforming them.

****

Trading Quality for Efficiency of Graph Partitioning: An Inductive Method across Graphs

Meng QIN,Chaorui Zhang,Bo Bai,Gong Zhang,Dit-Yan Yeung

Many applications of network systems can be formulated as several NP-hard combinatorial optimization problems regarding graph partitioning (GP), e.g., modularity maximization and NCut minimization. Due to the NP-hardness, to balance the quality and efficiency of GP remains a challenge. Existing methods use machine learning techniques to obtain high-quality solutions but usually have high complexity. Some fast GP methods adopt heuristic strategies to ensure low runtime but suffer from quality degradation. In contrast to conventional transductive GP methods applied to a static graph, we propose an inductive graph partitioning (IGP) framework across multiple evolving graph snapshots to alleviate the NP-hard challenge. IGP first conducts the offline training of a novel dual graph neural network on historical snapshots to capture the structural properties of a system. The trained model is then generalized to newly generated snapshots for fast high-quality online GP without additional optimization, where a better trade-off between quality and efficiency is achieved. IGP is also a generic framework that can capture the permutation invariant partitioning ground-truth of historical snapshot s in the offline training and tackle the online GP on graphs with non-fixed number of nodes and clusters. Experiments on a set of benchmarks demonstrate that IGP achieves competitive quality and efficiency to various state-of-the-art baselines.

****

Designing Less Forgetful Networks for Continual Learning

Nicholas I-Hsien Kuo,Mehrtash Harandi,Nicolas Fourrier,Gabriela Ferraro,Christian Walder,Hanna Suominen

Neural networks usually excel in learning a single task. Their weights are plastic and help them to learn quickly, but these weights are also known to be unstable. Hence, they may experience catastrophic forgetting and lose the ability to solve past tasks when assimilating information to solve a new task. Existing methods have mostly attempted to address this problem through external constraints. Replay shows the backbone network externally stored memories; regularisation imposes additional learning objectives; and dynamic architecture often introduces more parameters to host new knowledge. In contrast, we look for internal means to create less forgetful networks. This paper demonstrates that two simple architectural modifications -- Masked Highway Connection and Layer-Wise Normalisation -- can drastically reduce the forgetfulness in a backbone network. When naively employed to sequentially learn over multiple tasks, our modified backbones were as competitive as those unmodified backbones with continual learning techniques applied. Furthermore, our proposed architectural modifications were compatible with most if not all continual learning archetypes and therefore helped those respective techniques in achieving new state of the art.

****

Causal Discovery via Cholesky Factorization

Xu Li,YUNFENG CAI,Mingming Sun,Ping Li

Discovering the causal relationship via recovering the directed acyclic graph (DAG) structure from the observed data is a challenging combinatorial problem. This paper proposes an extremely fast, easy to implement, and high-performance DAG structure recovering algorithm. The algorithm is based on the Cholesky factorization of the covariance/precision matrix. The time complexity of the algorithm is $\mathcal{O}(p^2n + p^3)$, where p and n are the numbers of nodes and samples, respectively. Under proper assumptions, we show that our algorithm takes $\mathcal{O}(\log(p/\epsilon))$ samples to exactly recover the DAG structure with probability at least $1-\epsilon$. In both time and sample complexities, our algorithm is better than previous algorithms. On synthetic and real-world data sets, our algorithm is significantly faster than previous methods and achieves state-of-the-art performance.

****

Transferable Adversarial Attack based on Integrated Gradients

Yi Huang,Adams Wai-Kin Kong

The vulnerability of deep neural networks to adversarial examples has drawn tremendous attention from the community. Three approaches, optimizing standard objective functions, exploiting attention maps, and smoothing decision surfaces, are commonly used to craft adversarial examples. By tightly integrating the three approaches, we propose a new and simple algorithm named Transferable Attack based on Integrated Gradients (TAIG) in this paper, which can find highly transferable adversarial examples for black-box attacks. Unlike previous methods using multiple computational terms or combining with other methods, TAIG integrates the three approaches into one single term. Two versions of TAIG that compute their integrated gradients on a straight-line path and a random piecewise linear path are studied. Both versions offer strong transferability and can seamlessly work together with the previous methods. Experimental results demonstrate that TAIG outperforms the state-of-the-art methods.

****

How to deal with missing data in supervised deep learning?

Niels Bruun Ipsen,Pierre-Alexandre Mattei,Jes Frellsen

The issue of missing data in supervised learning has been largely overlooked, especially in the deep learning community. We investigate strategies to adapt neur al architectures for handling missing values. Here, we focus on regression and classification problems where the features are assumed to be missing at random. Of particular interest are schemes that allow reusing as-is a neural discriminative architecture. To address supervised deep learning with missing values, we propose to marginalize over missing values in a joint model of covariates and outcomes. Thereby, we leverage both the flexibility of deep generative models to describe the distribution of the covariates and the power of purely discriminative models to make predictions. More precisely, a deep latent variable model can be learned jointly with the discriminative model, using importance-weighted variational inference, essentially using importance sampling to mimick averaging over multiple imputations. In low-capacity regimes, or when the discriminative model has a strong inductive bias, we find that our hybrid generative/discriminative approach generally outperforms single imputations methods.

****

PMIC: Improving Multi-Agent Reinforcement Learning with Progressive Mutual Information Collaboration

Pengyi Li,Hongyao Tang,Tianpei Yang,Xiaotian Hao,Sang Tong,YAN ZHENG,Jianye HAO,Matthew E. Taylor,Jinyi Liu

Learning to collaborate is critical in multi-agent reinforcement learning (MARL). A branch of previous works proposes to promote collaboration by maximizing the correlation of agents’ behaviors, which is typically characterised by mutual information (MI) in different forms. However, simply maximizing the MI of agents’ behaviors cannot guarantee achieving better collaboration because suboptimal collaboration can also lead to high MI. In this paper, we first propose a new collaboration criterion to evaluate collaboration from three perspectives, which arrives at a form of the mutual information between global state and joint policy. This bypasses the introduction of explicit additional input of policies and mitigates the scalability issue meanwhile. Moreover, to better leverage MI-based collaboration signals, we propose a novel MARL framework, called Progressive Mutual Information Collaboration (PMIC) which contains two main components. The first component is Dual Progressive Collaboration Buffer (DPCB) which separately stores superior and inferior trajectories in a progressive manner. The second component is Dual Mutual Information Estimator (DMIE), including two neural estimators of our new designed MI based on separate samples in DPCB. We then make use of the neural MI estimates to improve agents' policies: to maximize the MI lower bound associated with superior collaboration to facilitate better collaboration and to minimize the MI upper bound associated with inferior collaboration to avoid falling into local optimal. PMIC is general and can be combined with existing MARL algorithms. Experiments on a wide range of MARL benchmarks show the superior performance of PMIC compared with other MARL algorithms.

****

Relative Entropy Gradient Sampler for Unnormalized Distributions

Xingdong Feng,Yuan Gao,Jian Huang,Yuling Jiao,Xu Liu

We propose a relative entropy gradient sampler (REGS) for sampling from unnormalized distributions. REGS is a particle method that seeks a sequence of simple nonlinear transforms iteratively pushing the initial samples from a reference distribution into the samples from an unnormalized target distribution. To determine the nonlinear transforms at each iteration, we consider the Wasserstein gradient flow of relative entropy. This gradient flow determines a path of probability distributions that interpolates the reference distribution and the target distribution. It is characterized by an ODE system with velocity fields depending on the density ratios of the density of evolving particles and the unnormalized target density. To sample with REGS, we need to estimate the density ratios and simulate the ODE system with particle evolution. We propose a novel nonparametric approach to estimating the logarithmic density ratio using neural networks. Extensive simulation studies on challenging multimodal 1D and 2D distributions and Bayesian logistic regression on real datasets demonstrate that the REGS outperforms the state-of-the-art sampling methods included in the comparison.

****

Provably Improved Context-Based Offline Meta-RL with Attention and Contrastive L earning

Lanqing Li,Yuanhao HUANG,Mingzhe Chen,siteng luo,Dijun Luo,Junzhou Huang

Meta-learning for offline reinforcement learning (OMRL) is an understudied problem with tremendous potential impact by enabling RL algorithms in many real-world applications. A popular solution to the problem is to infer task identity as augmented state using a context-based encoder, for which efficient learning of robust task representations remains an open challenge. In this work, we provably improve upon one of the SOTA OMRL algorithms, FOCAL, by incorporating intra-task attention mechanism and inter-task contrastive learning objectives, to robustify task representation learning against sparse reward and distribution shift. Theoretical analysis and experiments are presented to demonstrate the superior performance and robustness of our end-to-end and model-free framework compared to prior algorithms across multiple meta-RL benchmarks.

****

TransSlowDown: Efficiency Attacks on Neural Machine Translation Systems

Simin Chen,Mirazul Haque,Zihe Song,Cong Liu,Wei Yang

Neural machine translation (NMT) systems have received massive attention from academia and industry. Despite a rich set of work focusing on improving NMT systems’ accuracy, the less explored topic of efficiency is also important to NMT systems because of the real-time demand of translation applications. In this paper, we observe an inherent property of the NMT system, that is, NMT systems’ efficiency is related to the output length instead of the input length. Such property results in a new attack surface of the NMT system—an adversary can slightly changing inputs to incur a significant amount of redundant computations in NMT systems.   Such abuse of NMT systems’ computational resources is analogous to denial-of-service attacks. Abuse of NMT systems’ computing resources will affect the service quality (e.g., prolong response to users’ translation requests) and even make the translation service unavailable (e.g., running out of resources such as batteries of mobile devices). To further the understanding of such efficiency-oriented threats and raise the community’s concern on the efficiency robustness of NMT systems, we propose a new attack approach, TranSlowDown, to test the efficiency robustness of NMT systems. To demonstrate the effectiveness of TranSlowDown, we conduct a systematic evaluation on three public-available NMT systems: Google T5, Facebook Fairseq, and Helsinki-NLP translator. The experimental results show that TranSlowDown increases NMT systems’ response latency up to 1232%and 1056% on Intel CPU and Nvidia GPU respectively by inserting only three characters into existing input sentences. Our results also show that the adversarial examples generated byTranSlowDowncan consume more than 30 times battery power than the original benign example. Such results suggest that further research is required for protecting NMT systems against efficiency-oriented threats.

****

Context-Aware Sparse Deep Coordination Graphs

Tonghan Wang,Liang Zeng,Weijun Dong,Qianlan Yang,Yang Yu,Chongjie Zhang

Learning sparse coordination graphs adaptive to the coordination dynamics among agents is a long-standing problem in cooperative multi-agent learning. This paper studies this problem and proposes a novel method using the variance of payoff functions to construct context-aware sparse coordination topologies. We theoretically consolidate our method by proving that the smaller the variance of payoff functions is, the less likely action selection will change after removing the corresponding edge. Moreover, we propose to learn action representations to effectively reduce the influence of payoff functions' estimation errors on graph construction. To empirically evaluate our method, we present the Multi-Agent COordination (MACO) benchmark by collecting classic coordination problems in the literature, increasing their difficulty, and classifying them into different types. We carry out a case study and experiments on the MACO and StarCraft II micromanagement benchmark to demonstrate the dynamics of sparse graph learning, the influence of graph sparseness, and the learning performance of our method.

****

Genetic Algorithm for Constrained Molecular Inverse Design

Yurim Lee,Kyudam Choi,Cheongwon Kim

A genetic algorithm is suitable for exploring large search spaces as it finds an approximate solution. Because of this advantage, genetic algorithm is effective in exploring vast and unknown space such as molecular search space. Though the algorithm is suitable for searching vast chemical space, it is difficult to optimize pharmacological properties while maintaining molecular substructure. To solve this issue, we introduce a genetic algorithm featuring a constrained molecular inverse design. The proposed algorithm successfully produces valid molecules for crossover and mutation. Furthermore, it optimizes specific properties while adhering to structural constraints using a two-phase optimization. Experiments prove that our algorithm effectively finds molecules that satisfy specific properties while maintaining structural constraints.
****
On the approximation properties of recurrent encoder-decoder architectures
Zhong Li,Haotian Jiang,Qianxiao Li
Encoder-decoder architectures have recently gained popularity in sequence to sequence modelling, featuring in state-of-the-art models such as transformers. However, a mathematical understanding of their working principles still remains limited. In this paper, we study the approximation properties of recurrent encoder-decoder architectures. Prior work established theoretical results for RNNs in the linear setting, where approximation capabilities can be related to smoothness and memory of target temporal relationships. Here, we uncover that the encoder and decoder together form a particular “temporal product structure” which determines the approximation efficiency. Moreover, the encoder-decoder architecture generalises RNNs with the capability to learn time-inhomogeneous relationships. Our results provide the theoretical understanding of approximation properties of the recurrent encoder-decoder architecture, which precisely characterises, in the considered setting, the types of temporal relationships that can be efficiently learned.
****
NAIL: A Challenging Benchmark for Na\"ive Logical Reasoning
Xinbo Zhang,Changzhi Sun,Yue Zhang,Lei Li,Hao Zhou
Logical reasoning over natural text is an important capability towards human level intelligence.
Existing datasets are either limited and inadequate to train and evaluate logical reasoning capability (e.g., LogiQA and ReClor),
or not oriented for logical reasoning (e.g., SQuAD and HotpotQA).
In this paper, we focus on a specific category of logical reasoning, named \emph{\mytask}, and propose a new large scale benchmark, named \mydata, targeted for learning and evaluating models' capabilities towards \mytask.
\mydata is source from standardized exams such as Chinese National Civil Servants Examination and Law School Admission Test.
Furthermore, to collect more data, we propose to imitate the example of standardized exams rather than designing them from scratch.
\mydata is available in both Chinese and English containing a total of $10,296 * 2$ instances.
Empirical results show that current state-of-the-art neural models struggle on \mydata with very poor accuracy (the best result is 30.10\% for \mydata and 36.15\% for Chinese \mydata), while human experts can perform nearly 100\% accuracy.
Further results indicate that human imitations can significantly help models learn logic from natural text.
****
Structured Energy Network as a dynamic loss function. Case study. A case study with multi-label Classification
Jay-Yoon Lee,Dhruvesh Patel,Purujit Goyal,Andrew McCallum
We propose SEAL which utilizes this energy network as a trainable loss function for a simple feedfoward network. Structured prediction energy networks (SPENs) (Belanger & McCallum, 2016; Gygli et al., 2017) have shown that a neural network (i.e. energy network) can learn a reasonable energy function over the candidate structured outputs. We find that rather than using SPEN as a prediction network, using it as a trainable loss function is not only computationally efficient but

also results in higher performance. compared to SPENs in both training and inference time. As the energy loss function is trainable, we propose SEAL to be dynamic which can adapt energy function to focus on the region where feedforward model will be affected most. We find this to be effective in ablation study comparing SEAL to the static version (§4) where energy function is fixed after pretraining. We show the relation to previous work on the joint optimization model of energy network and feedforward model (INFNET) as we show that it is equivalent to SEAL using margin-based loss if INFNET relaxes their loss function. Based on the unique architecture of SEAL, we further propose a variant of SEAL that utilizes noise contrastive ranking (NCE) loss that by itself does not perform well as a structured energy network, but embodied in SEAL, it shows the greatest performance among the variants we study. We demonstrate the effectiveness of SEAL on 7 feature-based and 3 text-based multi-label classification datasets. The best version of SEAL that uses NCE ranking method achieves close to +2.85, +2.23 respective F1 point gain in average over cross-entropy and INFNET on the feature-based datasets, excluding one outlier that has an excessive gain of +50.0 F1 points. Lastly, examining whether the proposed framework is effective on a large pre-trained model as well, we observe SEAL achieving +0.87 F1 point gain in average on top of BERT-based adapter model o text datasets.
****

Closed-loop Control for Online Continual Learning
Yaqian Zhang,Eibe Frank,Bernhard Pfahringer,Albert Bifet,Nick Jin Sean Lim,Alvin Jia
Online class-incremental continual learning (CL) deals with the sequential task learning problem in a realistic non-stationary setting with a single-pass through of data. Replay-based CL methods have shown promising results in several online class-incremental continual learning benchmarks. However, these replay methods typically assume pre-defined and fixed replay dynamics, which is suboptimal. This paper introduces a closed-loop continual learning framework, which obtains a real-time feedback learning signal via an additional test memory and then adapts the replay dynamics accordingly. More specifically, we propose a reinforcement learning-based method to dynamically adjust replay hyperparameters online to balance the stability and plasticity trade-off in continual learning. To address the non-stationarity in the continual learning environment, we employ a Q function with task-specific and task-shared components to support fast adaptation. The proposed method is applied to improve state-of-the-art replay-based methods and achieves superior performance on popular benchmarks.
****

Pixelated Butterfly: Simple and Efficient Sparse training for Neural Network Models
Beidi Chen,Tri Dao,Kaizhao Liang,Jiaming Yang,Zhao Song,Atri Rudra,Christopher Re
Overparameterized neural networks generalize well but are expensive to train. Ideally one would like to reduce their computational cost while retaining their generalization benefits. Sparse model training is a simple and promising approach to achieve this, but there remain challenges as existing methods struggle with accuracy loss, slow training runtime, or difficulty in sparsifying all model components. The core problem is that searching for a sparsity mask over a discrete set of sparse matrices is difficult and expensive. To address this, our main insight is to optimize over a continuous superset of sparse matrices with a fixed structure known as products of butterfly matrices. As butterfly matrices are not hardware efficient, we propose simple variants of butterfly (block and flat) to take advantage of modern hardware. Our method (Pixelated Butterfly) uses a simple fixed sparsity pattern based on flat block butterfly and low-rank matrices to sparsify most network layers (e.g., attention, MLP). We empirically validate that Pixelated Butterfly is $3\times$ faster than Butterfly and speeds up training to achieve favorable accuracy--efficiency tradeoffs. On the ImageNet classification and WikiText-103 language modeling tasks, our sparse models train up to 2.3$\times$ faster than the dense MLP-Mixer, Vision Transformer, and GPT-2 small with no drop in accuracy.

****

Distributional Generalization: Structure Beyond Test Error
Preetum Nakkiran,Yamini Bansal

Classifiers in machine learning are often reduced to single dimensional quantities, such as test error or loss. Here, we initiate a much richer study of classifiers by considering the entire joint distribution of their inputs and outputs. We present both new empirical behaviors of standard classifiers, as well as quantitative conjectures which capture these behaviors. Informally, our conjecture states: the output distribution of an interpolating classifier matches the distribution of true labels, when conditioned on certain subgroups of the input space. For example, if we mislabel 30% of dogs as cats in the train set of CIFAR-10, then a ResNet trained to interpolation will in fact mislabel roughly 30% of dogs as cats on the *test set* as well, while leaving other classes unaffected. This conjecture has implications for the theory of overparameterization, scaling limits, implicit bias, and statistical consistency. Further, it can be seen as a new kind of generalization, which goes beyond measuring single-dimensional quantities to measuring entire distributions.

****

Topological Graph Neural Networks
Max Horn,Edward De Brouwer,Michael Moor,Yves Moreau,Bastian Rieck,Karsten Borgwardt

Graph neural networks (GNNs) are a powerful architecture for tackling graph learning tasks, yet have been shown to be oblivious to eminent substructures such as cycles. We present TOGL, a novel layer that incorporates global topological information of a graph using persistent homology. TOGL can be easily integrated into any type of GNN and is strictly more expressive (in terms the Weisfeiler–Lehman graph isomorphism test) than message-passing GNNs. Augmenting GNNs with TOGL leads to improved predictive performance for graph and node classification tasks, both on synthetic data sets, which can be classified by humans using their topology but not by ordinary GNNs, and on real-world data.

****

Learning Value Functions from Undirected State-only Experience
Matthew Chang,Arjun Gupta,Saurabh Gupta

This paper tackles the problem of learning value functions from undirected state-only experience (state transitions without action labels i.e. (s,s',r) tuples). We first theoretically characterize the applicability of Q-learning in this setting. We show that tabular Q-learning in discrete Markov decision processes (MDPs) learns the same value function under any arbitrary refinement of the action space. This theoretical result motivates the design of Latent Action Q-learning or LAQ, an offline RL method that can learn effective value functions from state-only experience. Latent Action Q-learning (LAQ) learns value functions using Q-learning on discrete latent actions obtained through a latent-variable future prediction model. We show that LAQ can recover value functions that have high correlation with value functions learned using ground truth actions. Value functions learned using LAQ lead to sample efficient acquisition of goal-directed behavior, can be used with domain-specific low-level controllers, and facilitate transfer across embodiments. Our experiments in 5 environments ranging from 2D grid world to 3D visual navigation in realistic environments demonstrate the benefits of LAQ over simpler alternatives, imitation learning oracles, and competing methods.

****

Generalizing MLPs With Dropouts, Batch Normalization, and Skip Connections
Taewoon Kim

A multilayer perceptron (MLP) is typically made of multiple fully connected layers with nonlinear activation functions. There have been several approaches to make them better (e.g., faster convergence, better convergence limit, etc.). But the researches lack structured ways to test them. We test different MLP architectures by carrying out the experiments on the age and gender datasets. We empirically show that by whitening inputs before every linear layer and adding skip connections, our proposed MLP architecture can result in better performance. Since t he whitening process includes dropouts, it can also be used to approximate Bayesian inference. We have open sourced our code, and released models and docker images at https://github.com/anonymous.

****

Diffusion-Based Voice Conversion with Fast Maximum Likelihood Sampling Scheme

Vadim Popov,Ivan Vovk,Vladimir Gogoryan,Tasnima Sadekova,Mikhail Sergeevich Kudinov,Jiansheng Wei

Voice conversion is a common speech synthesis task which can be solved in different ways depending on a particular real-world scenario. The most challenging one often referred to as one-shot many-to-many voice conversion consists in copying target voice from only one reference utterance in the most general case when both source and target speakers do not belong to the training dataset. We present a scalable high-quality solution based on diffusion probabilistic modeling and demonstrate its superior quality compared to state-of-the-art one-shot voice conversion approaches. Moreover, focusing on real-time applications, we investigate general principles which can make diffusion models faster while keeping synthesis quality at a high level. As a result, we develop a novel Stochastic Differential Equations solver suitable for various diffusion model types and generative tasks as shown through empirical studies and justify it by theoretical analysis.

****

Recurrent Parameter Generators

Jiayun Wang,Yubei Chen,Stella Yu,Brian Cheung,Yann LeCun

We present a generic method for recurrently using the same parameters for many different convolution layers to build a deep network. Specifically, for a network, we create a recurrent parameter generator (RPG), from which the parameters of each convolution layer are generated. Though using recurrent models to build a deep convolutional neural network (CNN) is not entirely new, our method achieves significant performance gain compared to the existing works. We demonstrate how to build a one-layer-size neural network to achieve similar performance compared to other traditional CNN models on various applications and datasets. We use the RPG to build a ResNet18 network with the number of weights equivalent to one convolutional layer of a conventional ResNet and show this model can achieve 67.2% ImageNet top-1 accuracy. Additionally, such a method allows us to build an arbitrarily complex neural network with any amount of parameters. For example, we build a ResNet34 with model parameters reduced by more than 400 times, which still achieves 41.6% ImageNet top-1 accuracy. Furthermore, the RPG can be further pruned and quantized for better run-time performance in addition to the model size reduction. We provide a new perspective for model compression. Rather than shrinking parameters from a large model, RPG sets a certain parameter-size constraint and uses the gradient descent algorithm to automatically find the best model under the constraint. Extensive experiment results are provided to demonstrate the power of the proposed recurrent parameter generator.

****

MANDERA: Malicious Node Detection in Federated Learning via Ranking

Wanchuang Zhu,Benjamin Zi Hao Zhao,Simon Luo,Ke Deng

Federated learning is a distributed learning paradigm which seeks to preserve the privacy of each participating node's data. However, federated learning is vulnerable to attacks, specifically to our interest, model integrity attacks. In this paper, we propose a novel method for malicious node detection called MANDERA. By transferring the original message matrix assembling the update gradients from local nodes into a ranking matrix that encodes the relative rankings of the outputs of all local nodes in different parameter dimensions, MANDERA seeks to distinguish the malicious nodes from the benign ones with high efficiency based on key characteristics of the rank domain. We have proved, under mild conditions, that MANDERA is guaranteed to detect all malicious nodes under typical Byzantine attacks with no prior knowledge or history about the participating nodes. The effectiveness of MANDERA is further confirmed by experiments on two classic datasets, CIFAR-10 and MNIST. Compared to the state-of-art methods in the literature for defending Byzantine attacks, MANDERA is unique in its way to identify the mal icious nodes by ranking and its robustness to effectively defense a wide range of attacks.

****

Privacy-preserving Task-Agnostic Vision Transformer for Image Processing

Boah Kim,Jeongsol Kim,Jong Chul Ye

Distributed collaborative learning approaches such as federated and split learning have attracted significant attention lately due to their ability to train neural networks using data from multiple sources without sharing data. However, they are not usually suitable in applications where each client carries out different tasks with its own data. Inspired by the recent success of Vision Transformer (ViT), here we present a new distributed learning framework for image processing applications, allowing clients to learn multiple tasks with their private data. The key idea arises from a novel task-agnostic Vision Transformer that is introduced to learn the global attention independent of specific tasks. Specifically, by connecting task-specific heads and tails at client sides to a task-agnostic Transformer body at a server side, each client learns a translation from its own task to a common representation, while the Transformer body learns global attention between the features embedded in the common representation. To enable decomposition between the task-specific and common representation, we propose an alternating training strategy in which task-specific learning for the heads and tails is run on the clients by fixing the Transformer, which alternates with task-agnostic learning for the Transformer on the server by freezing the heads and tails. Experimental results on multi-task learning for various image processing show that our method synergistically improves the performance of the task-specific network of each client while maintaining privacy.

****

Neural tangent kernel eigenvalues accurately predict generalization

James B Simon,Madeline Dickens,Michael Deweese

Finding a quantitative theory of neural network generalization has long been a central goal of deep learning research. We extend recent results to demonstrate that, by examining the eigensystem of a neural network's "neural tangent kernel," one can predict its generalization performance when learning arbitrary functions. Our theory accurately predicts not only test mean-squared-error but all first- and second-order statistics of the network's learned function. Furthermore, using a measure quantifying the "learnability" of a given target function, we prove a new "no free lunch" theorem characterizing a fundamental tradeoff in the inductive bias of wide neural networks: improving a network’s generalization for a given target function must worsen its generalization for orthogonal functions. We further demonstrate the utility of our theory by analytically predicting two surprising phenomena --- worse-than-chance generalization on hard-to-learn functions and nonmonotonic error curves in the small data regime --- which we subsequently observe in experiments. Though our theory is derived for infinite-width architectures, we find it agrees with networks as narrow as width 20, suggesting it is predictive of generalization in practical neural networks.

****

On the Impact of Client Sampling on Federated Learning Convergence

Yann Fraboni,Richard Vidal,Laetitia Kameni,Marco Lorenzi

While clients' sampling is a central operation of current state-of-the-art federated learning (FL) approaches, the impact of this procedure on the convergence and speed of FL remains under-investigated.In this work we introduce a novel decomposition theorem for the convergence of FL, allowing to clearly quantify the impact of client sampling on the global model update. Contrarily to previous convergence analyses, our theorem provides the exact decomposition of a given convergence step, thus enabling accurate considerations about the role of client sampling and heterogeneity. First, we provide a theoretical ground for previously reported experimental results on the relationship between FL convergence and the variance of the aggregation weights. Second, we prove for the first time that the quality of FL convergence is also impacted by the resulting \emph{covariance} between aggregation weights. Our theory is general, and is here applied to Multinomial Distribution (MD) and Uniform sampling, the two default client sampling sche mes of FL, and demonstrated through a series of experiments in non-iid and unbalanced scenarios. Our results suggest that MD sampling should be used as default sampling scheme, due to the resilience to the changes in data ratio during the learning process, while Uniform sampling is superior only in the special case when clients have the same amount of data.

****

Improving Fairness via Federated Learning

Yuchen Zeng,Hongxu Chen,Kangwook Lee

Recently, lots of algorithms have been proposed for learning a fair classifier from centralized data. However, how to privately train a fair classifier on decentralized data has not been fully studied yet. In this work, we first propose a new theoretical framework, with which we analyze the value of federated learning in improving fairness. Our analysis reveals that federated learning can strictly boost model fairness compared with all non-federated algorithms. We then theoretically and empirically show that the performance tradeoff of FedAvg-based fair learning algorithms is strictly worse than that of a fair classifier trained on centralized data. To resolve this, we propose FedFB, a private fair learning algorithm on decentralized data with a modified FedAvg protocol. Our extensive experimental results show that FedFB significantly outperforms existing approaches, sometimes achieving a similar tradeoff as the one trained on centralized data.

****

CPT: Colorful Prompt Tuning for Pre-trained Vision-Language Models

Yuan Yao,Ao Zhang,Zhengyan Zhang,Zhiyuan Liu,Tat-Seng Chua,Maosong Sun

Pre-Trained Vision-Language Models (VL-PTMs) have shown promising capabilities in grounding natural language in image data, facilitating a broad variety of cross-modal tasks. However, we note that there exists a significant gap between the objective forms of model pre-training and fine-tuning, resulting in a need for large amounts of labeled data to stimulate the visual grounding capability of VL-PTMs for downstream tasks. To address the challenge, we present Cross-modal Prompt Tuning (CPT, alternatively, Colorful Prompt Tuning), a novel paradigm for tuning VL-PTMs, which reformulates visual grounding into a fill-in-the-blank problem with color-based co-referential markers in image and text, maximally mitigating the gap. In this way, CPT enables strong few-shot and even zero-shot visual grounding capabilities of VL-PTMs. Comprehensive experimental results show that the prompt-tuned VL-PTMs outperform their fine-tuned counterparts by a large margin (e.g., 17.3% absolute accuracy improvement, and 73.8% relative standard deviation reduction on average with one shot in RefCOCO evaluation). All the data and codes will be available to facilitate future research.

****

The Boltzmann Policy Distribution: Accounting for Systematic Suboptimality in Human Models

Cassidy Laidlaw,Anca Dragan

Models of human behavior for prediction and collaboration tend to fall into two categories: ones that learn from large amounts of data via imitation learning, and ones that assume human behavior to be noisily-optimal for some reward function. The former are very useful, but only when it is possible to gather a lot of human data in the target environment and distribution. The advantage of the latter type, which includes Boltzmann rationality, is the ability to make accurate predictions in new environments without extensive data when humans are actually close to optimal. However, these models fail when humans exhibit systematic suboptimality, i.e. when their deviations from optimal behavior are not independent, but instead consistent over time. Our key insight is that systematic suboptimality can be modeled by predicting policies, which couple action choices over time, instead of trajectories. We introduce the Boltzmann policy distribution (BPD), which serves as a prior over human policies and adapts via Bayesian inference to capture systematic deviations by observing human actions during a single episode. The BPD is difficult to compute and represent because policies lie in a high-dimensional continuous space, but we leverage tools from generative and sequence modeling to enable efficient sampling and inference. We show that the BPD enables prediction of human behavior and human-AI collaboration equally as well as imi tation learning-based human models while using far less data.

****

SLIM-QN: A Stochastic, Light, Momentumized Quasi-Newton Optimizer for Deep Neural Networks

Yue Niu,Zalan Fabian,Sunwoo Lee,Mahdi Soltanolkotabi,Salman Avestimehr

We propose SLIM-QN, a light stochastic quasi-Newton optimizer for training large-scale deep neural networks (DNNs).

SLIM-QN addresses two key barriers in existing second-order methods for large-scale DNNs: 1) the high computational cost of obtaining the Hessian matrix and its inverse in every iteration (e.g. KFAC); 2) convergence instability due to stochastic training (e.g. L-BFGS).

To tackle the first challenge,SLIM-QN directly approximates the Hessian inverse using past parameters and gradients, without explicitly constructing the Hessian matrix and then computing its inverse.

To achieve stable convergence, SLIM-QN introduces momentum in Hessian updates together with an adaptive damping mechanism.

We provide rigorous theoretical results on the convergence of SLIM-QN in a stochastic setting.

We also demonstrate that SLIM-QN has much less compute and memory overhead compared to existing second-order methods.

To better understand the limitations and benefits of SLIM-QN, we evaluate its performance on various datasets and network architectures.

For instance on large datasets such as ImageNet, we show that SLIM-QN achieves near optimal accuracy $1.5\times$ faster when compared with SGD ($1.36\times$ faster in wall-clock time) using the same compute resources.

We also show that SLIM-QN can readily be applied to other contemporary non-convolutional architectures such as Transformers.

****

Synthesising Audio Adversarial Examples for Automatic Speech Recognition

Xinghua Qu,pengfei wei,Mingyong Gao,Zhu Sun,Yew-Soon Ong,Zejun MA

Adversarial examples in automatic speech recognition (ASR) are naturally sounded by humans yet capable of fooling well trained ASR models to transcribe incorrectly. Existing audio adversarial examples are typically constructed by adding constrained perturbations on benign audio inputs. Such attacks are therefore generated with an audio dependent assumption. For the first time, we propose the Speech Synthesising based Attack (SSA), a novel threat model that constructs audio adversarial examples entirely from scratch, i.e., without depending on any existing audio) to fool cutting-edge ASR models. To this end, we introduce a conditional variational auto-encoder (CVAE) as the speech synthesiser. Meanwhile, an adaptive sign gradient descent algorithm is proposed to solve the adversarial audio synthesis task. Experiments on three datasets (i.e., Audio Mnist, Common Voice, and Librispeech) show that our method could synthesise audio adversarial examples that are naturally sounded but misleading the start-of-the-art ASR models. The project webpage containing generated audio demos is at https://sites.google.com/view/ssa-asr/home.

****

WeakM3D: Towards Weakly Supervised Monocular 3D Object Detection

Liang Peng,Senbo Yan,Boxi Wu,Zheng Yang,Xiaofei He,Deng Cai

■Monocular 3D object detection is one of the most challenging tasks in 3D scene understanding. Due to the ill-posed nature of monocular imagery, existing monocular 3D detection methods highly rely on training with the manually annotated 3D box labels on the LiDAR point clouds. This annotation process is very laborious and expensive. To dispense with the reliance on 3D box labels, in this paper we explore the weakly supervised monocular 3D detection. Specifically, we first detect 2D boxes on the image. Then, we adopt the generated 2D boxes to select corresponding RoI LiDAR points as the weak supervision. Eventually, we adopt a network to predict 3D boxes which can tightly align with associated RoI LiDAR points. This network is learned by minimizing our newly-proposed 3D alignment loss between the 3D box estimates and the corresponding RoI LiDAR points. We will illustrate the potential challenges of the above learning problem and resolve these chal lenges by introducing several effective designs into our method. Codes are available at https://github.com/SPengLiang/WeakM3D.

****

Exploring Memorization in Adversarial Training

Yinpeng Dong,Ke Xu,Xiao Yang,Tianyu Pang,Zhijie Deng,Hang Su,Jun Zhu

Deep learning models have a propensity for fitting the entire training set even with random labels, which requires memorization of every training sample. In this paper, we explore the memorization effect in adversarial training (AT) for promoting a deeper understanding of model capacity, convergence, generalization, and especially robust overfitting of the adversarially trained models. We first demonstrate that deep networks have sufficient capacity to memorize adversarial examples of training data with completely random labels, but not all AT algorithms can converge under the extreme circumstance. Our study of AT with random labels motivates further analyses on the convergence and generalization of AT. We find that some AT approaches suffer from a gradient instability issue and the recently suggested complexity measures cannot explain robust generalization by considering models trained on random labels. Furthermore, we identify a significant drawback of memorization in AT that it could result in robust overfitting. We then propose a new mitigation algorithm motivated by detailed memorization analyses. Extensive experiments on various datasets validate the effectiveness of the proposed method.

****

8-bit Optimizers via Block-wise Quantization

Tim Dettmers,Mike Lewis,Sam Shleifer,Luke Zettlemoyer

Stateful optimizers maintain gradient statistics over time, e.g., the exponentially smoothed sum (SGD with momentum) or squared sum (Adam) of past gradient values. This state can be used to accelerate optimization significantly, compared to plain stochastic gradient descent, but uses memory that might otherwise be allocated to model parameters, thereby limiting the maximum size of models trained in practice. In this paper, we develop the first optimizers that use 8-bit statistics while maintaining the performance levels of using 32-bit optimizer states. To overcome the resulting computational, quantization, and stability challenges, we develop block-wise dynamic quantization. Block-wise quantization divides input tensors into smaller blocks that are independently quantized. Each block is processed in parallel across cores, yielding faster optimization and high precision quantization. To maintain stability and performance, we combine block-wise quantization with two additional changes: (1) dynamic quantization, a form of non-linear optimization that is precise for both large and small magnitude values, and (2) a stable embedding layer to reduce gradient variance that comes from the highly non-uniform distribution of input tokens in language models. As a result, our 8-bit optimizers maintain 32-bit performance with a small fraction of the memory footprint on a range of tasks, including 1.5B parameter language modeling, GLUE finetuning, ImageNet classification, WMT'14 machine translation, MoCo v2 contrastive ImageNet pretraining+finetuning, and RoBERTa pretraining, without changes to the original optimizer hyperparameters. We open-source our 8-bit optimizers as a drop-in replacement that only requires a two-line code change.

****

Decomposing Texture and Semantics for Out-of-distribution Detection

Jeong-Hyeon Moon,Namhyuk Ahn,Kyung-Ah Sohn

Out-of-distribution (OOD) detection has made significant progress in recent years because the distribution mismatch between the training and testing can severely deteriorate the reliability of a machine learning system.Nevertheless, the lack of precise interpretation of the in-distribution limits the application of OOD detection methods to real-world system pipielines. To tackle this issue, we decompose the definition of the in-distribution into texture and semantics, motivated by real-world scenarios. In addition, we design new benchmarks to measure the robustness that OOD detection methods should have. To achieve a good balance between the OOD detection performance and robustness, our method takes a divide-and-conquer approach. That is, the model first tackles each component of the textu re and semantics separately, and then combines them later. Such design philosophy is empirically proven by a series of benchmarks including not only ours but also the conventional counterpart.

****

Global Magnitude Pruning With Minimum Threshold Is All We Need
Manas Gupta,Vishandi Rudy Keneta,Abhishek Vaidyanathan,Ritwik Kanodia,Efe Camci,Chuan-Sheng Foo,Jie Lin

Neural network pruning remains a very important yet challenging problem to solve. Many pruning solutions have been proposed over the years with high degrees of algorithmic complexity. In this work, we shed light on a very simple pruning technique that achieves state-of-the-art (SOTA) performance. We showcase that magnitude based pruning, specifically, global magnitude pruning (GP) is sufficient to achieve SOTA performance on a range of neural network architectures. In certain architectures, the last few layers of a network may get over-pruned. For these cases, we introduce a straightforward method to mitigate this. We preserve a certain fixed number of weights in each layer of the network to ensure no layer is over-pruned. We call this the Minimum Threshold (MT). We find that GP combined with MT when needed, achieves SOTA performance on all datasets and architectures tested including ResNet-50 and MobileNet-V1 on ImageNet. Code available on github.

****

Disentanglement Analysis with Partial Information Decomposition
Seiya Tokui,Issei Sato

We propose a framework to analyze how multivariate representations disentangle ground-truth generative factors. A quantitative analysis of disentanglement has been based on metrics designed to compare how one variable explains each generative factor. Current metrics, however, may fail to detect entanglement that involves more than two variables, e.g., representations that duplicate and rotate generative factors in high dimensional spaces. In this work, we establish a framework to analyze information sharing in a multivariate representation with Partial Information Decomposition and propose a new disentanglement metric. This framework enables us to understand disentanglement in terms of uniqueness, redundancy, and synergy. We develop an experimental protocol to assess how increasingly entangled representations are evaluated with each metric and confirm that the proposed metric correctly responds to entanglement. Through experiments on variational autoencoders, we find that models with similar disentanglement scores have a variety of characteristics in entanglement, for each of which a distinct strategy may be required to obtain a disentangled representation.

****

Test-time Batch Statistics Calibration for Covariate Shift
fuming you,Jingjing Li,Zhou Zhao

Deep neural networks have a clear degradation when applying to the unseen environment due to the covariate shift. Conventional approaches like domain adaptation requires the pre-collected target data for iterative training, which is impractical in real-world applications. In this paper, we propose to adapt the deep models to the novel environment during inference. An previous solution is test time normalization, which substitutes the source statistics in BN layers with the target batch statistics. However, we show that test time normalization may potentially deteriorate the discriminative structures due to the mismatch between target batch statistics and source parameters. To this end, we present a general formulation α-BN to calibrate the batch statistics by mixing up the source and target statistics for both alleviating the domain shift and preserving the discriminative structures. Based on α-BN, we further present a novel loss function to form a unified test time adaptation framework Core, which performs the pairwise class correlation online optimization. Extensive experiments show that our approaches achieve the state-of-the-art performance on total twelve datasets from three topics, including model robustness to corruptions, domain generalization on image classification and semantic segmentation. Particularly, our α-BN improves 28.4\% to 43.9\% on GTA5 $\rightarrow$ Cityscapes without any training, even outperforms the latest source-free domain adaptation method.

****

A General Unified Graph Neural Network Framework Against Adversarial Attacks

Yujie Gu,Yangkun Cao,Qiang Huang,Huiyan Sun

Graph Neural Networks (GNNs) are powerful tools in representation learning for graphs. However, they are reported to be vulnerable to adversarial attacks, raising numerous concerns for applying it in some risk-sensitive domains. Therefore, it is essential to develop a robust GNN model to defend against adversarial attacks. Existing studies address this issue only considering cleaning perturbed graph structure, and almost none of them simultaneously consider denoising features. As the graph and features are interrelated and influence each other, we propose a General Unified Graph Neural Network (GUGNN) framework to jointly clean the graph and denoise features of data. On this basis, we further extend it by introducing two operations and develop a robust GNN model(R-GUGNN) to defend against adversarial attacks. One operation is reconstructing the graph with its intrinsic properties, including similarity of two adjacent nodes’ features, sparsity of real-world graphs and many slight noises having small eigenvalues in perturbed graphs. The other is the convolution operation for features to find the optimal solution adopting the Laplacian smoothness and the prior knowledge that nodes with many neighbors are difficult to attack. Experiments on four real-world datasets demonstrate that R-GUGNN has greatly improved the overall robustness over the state-of-the-art baselines.

****

Differentiable Gradient Sampling for Learning Implicit 3D Scene Reconstructions from a Single Image

Shizhan Zhu,Sayna Ebrahimi,Angjoo Kanazawa,Trevor Darrell

Implicit shape models are promising 3D representations for modeling arbitrary locations, with Signed Distance Functions (SDFs) particularly suitable for clear mesh surface reconstruction. Existing approaches for single object reconstruction impose supervision signals based on the loss of the signed distance value from all locations in a scene, posing difficulties when extending to real-world scenarios. The spatial gradient of the signed distance field, rather than the SDF value itself, has not been typically employed as a source of supervision for single-view reconstruction, in part due to the difficulties of differentiable sampling a spatial gradient from the feature map. In this study, we derive a novel closed-form gradient sampling solution for Differentialble Gradient Sampling (DGS) that enables backpropagation of the loss of the spatial gradient back to the feature map pixels, thus allowing the imposition of the loss efficiently on the spatial gradient. As a result, we achieve high-quality single view indoor scene reconstruction results learning directly from a real-world scanned dataset (e.g. ScannetV2). Our model also performs well when generalizing to unseen images downloaded directly from the internet (Fig. 1). We comfortably advanced the state-of-the-art results with several established datasets including ShapeNet and ScannetV2; extensive quantitative analysis confirmed that our proposed DGS module plays an essential role in achieving this performance improvement. Full codes are available in MaskedURL.

****

ContraQA: Question Answering under Contradicting Contexts

Liangming Pan,Wenhu Chen,Min-Yen Kan,William Yang Wang

With a rise in false, inaccurate, and misleading information in propaganda, news, and social media, real-world Question Answering (QA) systems face the challenges of synthesizing and reasoning over contradicting information to derive correct answers. This urgency gives rise to the need to make QA systems robust to misinformation, a topic previously unexplored. We study the risk of misinformation to QA models by investigating the behavior of the QA model under contradicting contexts that are mixed with both real and fake information. We create the first large-scale dataset for this problem, namely ContraQA, which contains over 10K human-written and model-generated contradicting pairs of contexts. Experiments show that QA models are vulnerable under contradicting contexts brought by misinformation. To defend against such a threat, we build a misinformation-aware QA system as a counter-measure that integrates question answering and misinformation de tection in a joint fashion.

****

MixRL: Data Mixing Augmentation for Regression using Reinforcement Learning

Seong-Hyeon Hwang,Steven Euijong Whang

Data augmentation is becoming essential for improving regression accuracy in critical applications including manufacturing and finance. Existing techniques for data augmentation largely focus on classification tasks and do not readily apply to regression tasks. In particular, the recent Mixup techniques for classification rely on the key assumption that linearity holds among training examples, which is reasonable if the label space is discrete, but has limitations when the label space is continuous as in regression. We show that mixing examples that either have a large data or label distance may have an increasingly-negative effect on model performance. Hence, we use the stricter assumption that linearity only holds within certain data or label distances for regression where the degree may vary by each example. We then propose MixRL, a data augmentation meta learning framework for regression that learns for each example how many nearest neighbors it should be mixed with for the best model performance using a small validation set. MixRL achieves these objectives using Monte Carlo policy gradient reinforcement learning. Our experiments conducted both on synthetic and real datasets show that MixRL significantly outperforms state-of-the-art data augmentation baselines. MixRL can also be integrated with other classification Mixup techniques for better results.

****

SubMix: Practical Private Prediction for Large-scale Language Models

Tony A Ginart,Laurens van der Maaten,James Zou,Chuan Guo

Recent data-extraction attacks have exposed that language models can memorize some training samples verbatim. This is a vulnerability that can compromise the privacy of the model’s training data. In this work, we introduce SubMix a practical protocol for private next-token prediction designed to prevent privacy violations by language models that were fine-tuned on a private corpus after pre-training on a public corpus. We show that SubMix limits the leakage of information that is unique to any individual user in the private corpus via a relaxation of group differentially private prediction. Importantly, SubMix admits a tight, data-dependent privacy accounting mechanism, which allows it to thwart existing data-extraction attacks while maintaining the utility of the language model. SubMix is the first protocol that maintains privacy even when publicly releasing tens of thousands of next-token predictions made by large transformer-based models such as GPT-2.

****

MDFL: A UNIFIED FRAMEWORK WITH META-DROPOUT FOR FEW-SHOT LEARNING

Shaobo Lin,Xingyu Zeng,Rui Zhao

Conventional training of deep neural networks usually requires a substantial amount of data with expensive human annotations. In this paper, we utilize the idea of meta-learning to integrate two very different streams of few-shot learning, i.e., the episodic meta-learning-based and pre-train finetune-based few-shot learning, and form a unified meta-learning framework. In order to improve the generalization power of our framework, we propose a simple yet effective strategy named meta-dropout, which is applied to the transferable knowledge generalized from base categories to novel categories. The proposed strategy can effectively prevent neural units from co-adapting excessively in the meta-training stage. Extensive experiments on the few-shot object detection and few-shot image classification datasets, i.e., Pascal VOC, MS COCO, CUB, and mini-ImageNet, validate the effectiveness of our method.

****

Neural Combinatorial Optimization with Reinforcement Learning : Solving theVehicle Routing Problem with Time Windows

Abdelhakim Abdellaoui,Issmail El Hallaoui,Loubna Benabbou

In contrast to the classical techniques for solving combinatorial optimization problems, recent advancements in reinforcement learning yield the potential to in dependently learn heuristics without any human interventions. In this context, the current paper aims to present a complete framework for solving the vehicle routing problem with time windows (VRPTW) relying on neural networks and reinforcement learning. Our approach is mainly based on an attention model (AM) that predicts the near-optimal distribution over different problem instances. To optimize its parameters, this model is trained in a reinforcement learning(RL) environment using a stochastic policy gradient and through a real-time evaluation of the reward, quantity to meet the problem business and logical constraints. Using synthetic data, the proposed model outperforms some existing baselines. This performance comparison was on the basis of the solution quality (total tour length) and the computation time (inference time) for small and medium-sized samples.

****

DRIBO: Robust Deep Reinforcement Learning via Multi-View Information Bottleneck
Jiameng Fan,Wenchao Li

Deep reinforcement learning (DRL) agents are often sensitive to visual changes that were unseen in their training environments. To address this problem, we leverage the sequential nature of RL to learn robust representations that encode only task-relevant information from observations based on the unsupervised multi-view setting. Specifically, we introduce a novel contrastive version of Multi-View Information Bottleneck (MIB) objective for temporal data. We train RL agents from pixels with this auxiliary objective to learn robust representations that can compress away task-irrelevant information and are predictive of task-relevant dynamics. This approach enables us to train high-performance policies that are robust to visual distractions and can generalize well to unseen environments. We demonstrate that our approach can achieve SOTA performance on diverse visual control tasks on the DeepMind Control Suite when the background is replaced with natural videos. In addition, we show that our approach outperforms well-established baselines for generalization to unseen environments on the Procgen benchmark.

****

Learning Continuous Environment Fields via Implicit Functions
Xueting Li,Shalini De Mello,Xiaolong Wang,Ming-Hsuan Yang,Jan Kautz,Sifei Liu

We propose a novel scene representation that encodes reaching distance -- the distance between any position in the scene to a goal along a feasible trajectory. We demonstrate that this environment field representation can directly guide the dynamic behaviors of agents in 2D mazes or 3D indoor scenes. Our environment field is a continuous representation and learned via a neural implicit function using discretely sampled training data. We showcase its application for agent navigation in 2D mazes, and human trajectory prediction in 3D indoor environments. To produce physically plausible and natural trajectories for humans, we additionally learn a generative model that predicts regions where humans commonly appear, and enforce the environment field to be defined within such regions. Extensive experiments demonstrate that the proposed method can generate both feasible and plausible trajectories efficiently and accurately.

****

LatentKeypointGAN: Controlling GANs via Latent Keypoints
Xingzhe He,Bastian Wandt,Helge Rhodin

Generative adversarial networks (GANs) have attained photo-realistic quality in image generation. However, how to best control the image content remains an open challenge. We introduce LatentKeypointGAN, a two-stage GAN which is trained end-to-end on the classical GAN objective with internal conditioning on a set of space keypoints. These keypoints have associated appearance embeddings that respectively control the position and style of the generated objects and their parts. A major difficulty that we address with suitable network architectures and training schemes is disentangling the image into spatial and appearance factors without domain knowledge and supervision signals. We demonstrate that LatentKeypointGAN provides an interpretable latent space that can be used to re-arrange the generated images by re-positioning and exchanging keypoint embeddings, such as generating portraits by combining the eyes, nose, and mouth from different images. In addition, the explicit generation of keypoints and matching images enables a new, GAN-based method for unsupervised keypoint detection.

****

C+1 Loss: Learn to Classify C Classes of Interest and the Background Class Differentially

Changhuai Chen,Xile Shen,Mengyu Ye,Yi Lu,Jun Che,Shiliang Pu

There is one kind of problem all around the classification area, where we want to classify C+1 classes of samples, including C semantically deterministic classes which we call classes of interest and the (C+1)th semantically undeterministic class which we call background class. In spite of most classification algorithm use softmax-based cross-entropy loss to supervise the classifier training process without differentiating the background class from the classes of interest, it is unreasonable as each of the classes of interest has its own inherent characteristics, but the background class dosen’t. We figure out that the background class should be treated differently from the classes of interest during training. Motivated by this, firstly we define the C+1 classification problem. Then, we propose three properties that a good C+1 classifier should have: basic discriminability, compactness and background margin. Based on them we define a uniform general C+1 loss, composed of three parts, driving the C+1 classifier to satisfy those properties. Finally, we instantialize a C+1 loss and experiment it in semantic segmentation, human parsing and object detection tasks. The proposed approach shows its superiority over the traditional cross-entropy loss.

****

Generalizable Learning to Optimize into Wide Valleys

Junjie Yang,Tianlong Chen,Mingkang Zhu,Fengxiang He,Dacheng Tao,Yingbin Liang,Zhangyang Wang

Learning to optimize (L2O) has gained increasing popularity in various optimization tasks, since classical optimizers usually require laborious, problem-specific design and hyperparameter tuning. However, current L2O approaches are designed for fast minimization of the objective function value (i.e., training error), hence often suffering from poor generalization ability such as in training deep neural networks (DNNs), including (i) disappointing performance across unseen optimizees $\textit{(optimizer generalization)}$; ($ii$) unsatisfactory test-set accuracy of trained DNNs ($\textit{optmizee generalization}$). To overcome the limitations, this paper introduces $\textit{flatness-aware}$ regularizers into L2O for shaping the local geometry of optimizee's loss landscape. Specifically, it guides optimizee to locate well-generalizable minimas in large flat regions of loss surface, while tending to avoid sharp valleys. Such optimizee generalization abilities of $\textit{flatness-aware}$ regularizers have been proved theoretically. Extensive experiments consistently validate the effectiveness of our proposals with substantially improved generalization on multiple sophisticated L2O models and diverse optimizees. Our theoretical and empirical results solidify the foundation for L2O's practically usage. All codes and pre-trained models will be shared upon acceptance.

****

Learning to Abstain in the Presence of Uninformative Data

Yikai Zhang,Songzhu Zheng,Pengxiang Wu,Yuriy Nevmyvaka,Chao Chen

Learning and decision making in domains with naturally high noise-to-signal ratios – such as Finance or Public Health – can be challenging and yet extremely important. In this paper, we study a problem of learning on datasets in which a significant proportion of samples does not contain useful information. To analyze this setting, we introduce a noisy generative process with a clear distinction between uninformative/not learnable/purely random data and a structured/informative component. This dichotomy is present both during the training and in the inference phase. We propose a novel approach to learn under these conditions via a loss inspired by the selective learning theory. By minimizing the loss, our method is guaranteed to make a near-optimal decision by simultaneously distinguishing structured data from the non-learnable and making predictions, even in a highly imbalanced setting. We build upon the strength of our theoretical guarantees by describing an iterative algorithm, which jointly optimizes both a predictor and a selector, and evaluate its empirical performance under a variety of conditions.

****

On the Convergence and Calibration of Deep Learning with Differential Privacy

Zhiqi Bu,Hua Wang,Qi Long,Weijie J Su

In deep learning with differential privacy (DP), the neural network achieves the privacy usually at the cost of slower convergence (and thus lower performance) than its non-private counterpart. This work gives the first convergence analysis of the DP deep learning, through the lens of training dynamics and the neural tangent kernel (NTK) matrix. Our convergence theory successfully characterizes the effects of two key components in the DP training: the per-sample clipping and the noise addition. We initiate a general principled framework to understand the DP deep learning with any network architecture, loss function and various optimizers including DP-Adam. Our analysis also motivates a new clipping method, the 'global clipping', that significantly improves the convergence, as well as preserves the same DP guarantee and computational efficiency as the existing method, which we term as 'local clipping'. In addition, our global clipping is surprisingly effective at learning calibrated classifiers, in contrast to the existing DP classifiers which are oftentimes over-confident and unreliable. Implementation-wise, the new clipping can be realized by inserting one line of code into the Pytorch Opacus library.

****

Hierarchical Cross Contrastive Learning of Visual Representations

Hesen Chen,Ming Lin,Xiuyu Sun,Rong Jin

The rapid progress of self-supervised learning (SSL) has greatly reduced the labeling cost in computer vision. The key idea of SSL is to learn invariant visual representations by maximizing the similarity between different views of the same input image. In most SSL methods, the representation invariant is measured by a contrastive loss which compares one of the network outputs after the projection head to its augmented version. Albeit being effective, this approach overlooks the information containing in the hidden layer of the projection head therefore could be sub-optimal. In this work, we propose a novel approach termed Hierarchical Cross Contrastive Learning(HCCL) to further distill the information mismatched by the conventional contrastive loss. The HCCL uses a hierarchical projection head to project the raw representations of the backbone into multiple latent spaces and then compares latent features across different levels and different views. By cross-level contrastive learning, HCCL not only regulates invariant on multiple hidden levels but also crosses different levels, improving the generalization ability of the learned visual representations. As a simple and generic method, HCCL can be applied to different SSL frameworks. We validate the efficacy of HCCL under classification, detection, segmentation, and few-shot learning tasks. Extensive experimental results show that HCCL outperforms most previous methods in various benchmark datasets.

****

Causal Contextual Bandits with Targeted Interventions

Chandrasekar Subramanian,Balaraman Ravindran

We study a contextual bandit setting where the learning agent has the ability to perform interventions on targeted subsets of the population, apart from possessing qualitative causal side-information. This novel formalism captures intricacies in real-world scenarios such as software product experimentation where targeted experiments can be conducted. However, this fundamentally changes the set of options that the agent has, compared to standard contextual bandit settings, necessitating new techniques. This is also the first work that integrates causal side-information in a contextual bandit setting, where the agent aims to learn a policy that maps contexts to arms (as opposed to just identifying one best arm). We propose a new algorithm, which we show empirically performs better than baselines on experiments that use purely synthetic data and on real world-inspired experiments. We also prove a bound on regret that theoretically guards performance.

****

Conditional GANs with Auxiliary Discriminative Classifier

Liang Hou,Qi Cao,Huawei Shen,Xueqi Cheng

Conditional generative models aim to learn the underlying joint distribution of data and labels, and thus realize conditional generation. Among them, auxiliary classifier generative adversarial networks (AC-GAN) have been widely used, but suffer from the problem of low intra-class diversity on generated samples. In this paper, we point out that the fundamental reason is that the classifier of AC-GAN is generator-agnostic, and therefore cannot provide informative guidance to the generator to approximate the target distribution, resulting in minimization of conditional entropy that decreases the intra-class diversity. Motivated by this observation, we propose a novel conditional GAN with auxiliary \textit{discriminative} classifier (ADC-GAN) to resolve the problem of AC-GAN. Specifically, the proposed auxiliary \textit{discriminative} classifier becomes generator-aware by recognizing the labels of the real data and the generated data \textit{discriminatively}. Our theoretical analysis reveals that the generator can faithfully replicate the target distribution even without the original discriminator, making the proposed ADC-GAN robust to the hyper-parameter and stable during the training process. Extensive experimental results on synthetic and real-world datasets demonstrate the superiority of ADC-GAN on conditional generative modeling compared to competing methods.

****

Finding Biological Plausibility for Adversarially Robust Features via Metameric Tasks

Anne Harrington,Arturo Deza

Recent work suggests that feature constraints in the training datasets of deep neural networks (DNNs) drive robustness to adversarial noise (Ilyas et al., 2019). The representations learned by such adversarially robust networks have also been shown to be more human perceptually-aligned than non-robust networks via image manipulations (Santurkar et al., 2019, Engstrom et al., 2019). Despite appearing closer to human visual perception, it is unclear if the constraints in robust DNN representations match biological constraints found in human vision. Human vision seems to rely on texture-based/summary statistic representations in the periphery, which have been shown to explain phenomena such as crowding (Balas et al., 2009) and performance on visual search tasks (Rosenholtz et al., 2012). To understand how adversarially robust optimizations/representations compare to human vision, we performed a psychophysics experiment using a metamer task similar to Freeman \& Simoncelli, 2011, Wallis et al., 2016 and Deza et al., 2019 where we evaluated how well human observers could distinguish between images synthesized to match adversarially robust representations compared to non-robust representations and a texture synthesis model of peripheral vision (Texforms a la Long et al., 2018). We found that the discriminability of robust representation and texture model images decreased to near chance performance as stimuli were presented farther in the periphery. Moreover, performance on robust and texture-model images showed similar trends within participants, while performance on non-robust representations changed minimally across the visual field. These results together suggest that (1) adversarially robust representations capture peripheral computation better than non-robust representations and (2) robust representations capture peripheral computation similar to current state-of-the-art texture peripheral vision models. More broadly, our findings support the idea that localized texture summary statistic representations may drive human invariance to adversarial perturbations and that the incorporation of such representations in DNNs could give rise to useful properties like adversarial robustness.

****

A Unified Knowledge Distillation Framework for Deep Directed Graphical Models

Yizhuo Chen,Kaizhao Liang,Zhe Zeng,Yifei Yang,Shuochao Yao,Huajie Shao

Knowledge distillation (KD) is a technique that transfers the knowledge from a large teacher network to a small student network. It has been widely applied to many different tasks, such as model compression and federated learning. However, the existing KD methods fail to generalize to general \textit{deep directed graphical models (DGMs)} with arbitrary layers of random variables. We refer by \textit{deep} DGMs to DGMs whose conditional distributions are parameterized by deep neural networks. In this work, we propose a novel unified knowledge distillati on framework for deep DGMs on various applications. Specifically, we leverage the reparameterization trick to hide the intermediate latent variables, resulting in a compact DGM. Then we develop a surrogate distillation loss to reduce error accumulation through multiple layers of random variables. Moreover, we present the connections between our method and some existing knowledge distillation approaches. The proposed framework is evaluated on three applications: deep generative models compression, discriminative deep DGMs compression, and VAE continual learning. The results show that our distillation method outperforms the baselines in data-free compression of deep generative models, including variational autoencoder (VAE), variational recurrent neural networks (VRNN), and Helmholtz Machine (HM). Moreover, our method achieves good performance for discriminative deep DGMs compression. Finally, we also demonstrate that it significantly improves the continual learning performance of VAE.

****

Post-Training Quantization Is All You Need to Perform Cross-Platform Learned Image Compression

Dailan He,Ziming Yang,Yan Wang,Yuan Chen,Qi Zhang,Hongwei Qin

It has been witnessed that learned image compression has outperformed conventional image coding techniques and tends to be practical in industrial applications. One of the most critical issues preventing it from being practical is the non-deterministic calculation, which makes the probability prediction cross-platform inconsistent and frustrates successful decoding. We propose to solve this problem by introducing well-developed post-training quantization and making the model inference integer-arithmetic-only, which is much simpler than presently existing training and fine-tuning based approaches yet still keeps the superior rate-distortion performance of learned image compression. Based on that, we further improve the discretization of the entropy parameters and extend the deterministic inference to fit Gaussian mixture models. With our proposed methods, the current state-of-the-art image compression models can infer in a cross-platform consistent manner, which makes the further development and practice of learned image compression more promising.

****

Lifting Imbalanced Regression with Self-Supervised Learning

Weiguo Pian,Hanyu Peng,Mingming Sun,Ping Li

A new influential task called imbalanced regression, most recently inspired by imbalanced classification, originating straightforwardly from both the imbalance and regression worlds, has received a great deal of attention. Yet we are still at a fairly preliminary stage in the exploration of this task, so more attempts are needed. In this paper, we work on a seamless marriage of imbalanced regression and self-supervised learning. But with this comes the first question of how to measure the similarity and dissimilarity under the regression sense, for which the definition is clear in the classification. To overcome the limitation, the formal definition of similarity in the regression task is given. On top of this, through experimenting on a simple neural network, we found that self-supervised learning could help alleviate the problem. However, the second problem is, it is not guaranteed that the noisy samples are similar to original samples when scaling to a deep network by adding random noise to the input, we specifically propose to limit the volume of noise on the output, and in doing so to find meaningful noise on the input by back propagation. Experimental results show that our approach achieves the state-of-the-art performance.

****

Sound and Complete Neural Network Repair with Minimality and Locality Guarantees

Feisi Fu,Wenchao Li

We present a novel methodology for repairing neural networks that use ReLU activation functions. Unlike existing methods that rely on modifying the weights of a neural network which can induce a global change in the function space, our approach applies only a localized change in the function space while still guaranteeing the removal of the buggy behavior. By leveraging the piecewise linear nature of ReLU networks, our approach can efficiently construct a patch network tailored to the linear region where the buggy input resides, which when combined with the original network, provably corrects the behavior on the buggy input. Our method is both sound and complete -- the repaired network is guaranteed to fix the buggy input, and a patch is guaranteed to be found for any buggy input. Moreover, our approach preserves the continuous piecewise linear nature of ReLU networks, automatically generalizes the repair to all the points including other undetected buggy inputs inside the repair region, is minimal in terms of changes in the function space, and guarantees that outputs on inputs away from the repair region are unaltered. On several benchmarks, we show that our approach significantly outperforms existing methods in terms of locality and limiting negative side effects.

****

Blaschke Product Neural Networks (BPNN): A Physics-Infused Neural Network for Phase Retrieval of Meromorphic Functions

Juncheng Dong,Simiao Ren,Yang Deng,Omar Khatib,Jordan Malof,Mohammadreza Soltani,Willie Padilla,Vahid Tarokh

Numerous physical systems are described by ordinary or partial differential equations whose solutions are given by holomorphic or meromorphic functions in the complex domain. In many cases, only the magnitude of these functions are observed on various points on the purely imaginary $j\omega$-axis since coherent measurement of their phases is often expensive. However, it is desirable to retrieve the lost phases from the magnitudes when possible. To this end, we propose a physics-infused deep neural network based on the Blaschke products for phase retrieval. Inspired by the Helson and Sarason Theorem, we recover coefficients of a rational function of Blaschke products using a Blaschke Product Neural Network (BPNN), based upon the magnitude observations as input. The resulting rational function is then used for phase retrieval. We compare the BPNN to conventional deep neural networks (NNs) on several phase retrieval problems, comprising both synthetic and contemporary real-world problems (e.g., metamaterials for which data collection requires substantial expertise and is time consuming). On each phase retrieval problem, we compare against a population of conventional NNs of varying size and hyperparameter settings. Even without any hyper-parameter search, we find that BPNNs consistently outperform the population of optimized NNs in scarce data scenarios, and do so despite being much smaller models. The results can in turn be applied to calculate the refractive index of metamaterials, which is an important problem in emerging areas of material science.

****

Variability of Neural Networks and Han-Layer: A Variability-Inspired Model

Yueyao Yu,Yin Zhang

What makes an artificial neural network easier to train or to generalize better than its peers? We introduce a notion of variability to view such issues under the setting of a fixed number of parameters which is, in general, a dominant cost-factor. Experiments verify that variability correlates positively to the number of activations and negatively to a phenomenon called Collapse to Constants, which is related but not identical to vanishing gradient. Further experiments on stylized problems show that variability is indeed a key performance indicator for fully-connected neural networks. Guided by variability considerations, we propose a new architecture called Householder-absolute neural layers, or Han-layers for short, to build high variability networks with a guaranteed immunity to gradient vanishing or exploding.

On small stylized models, Han-layer networks exhibit a far superior generalization ability over fully-connected networks. Extensive empirical results demonstrate that, by judiciously replacing fully-connected layers in large-scale networks such as MLP-Mixers, Han-layers can greatly reduce the number of model parameters while maintaining or improving generalization performance. We will also briefly discuss current limitations of the proposed Han-layer architecture.

****

Automated Self-Supervised Learning for Graphs

Wei Jin,Xiaorui Liu,Xiangyu Zhao,Yao Ma,Neil Shah,Jiliang Tang

Graph self-supervised learning has gained increasing attention due to its capaci ty to learn expressive node representations. Many pretext tasks, or loss functions have been designed from distinct perspectives. However, we observe that different pretext tasks affect downstream tasks differently cross datasets, which suggests that searching pretext tasks is crucial for graph self-supervised learning. Different from existing works focusing on designing single pretext tasks, this work aims to investigate how to automatically leverage multiple pretext tasks effectively. Nevertheless, evaluating representations derived from multiple pretext tasks without direct access to ground truth labels makes this problem challenging. To address this obstacle, we make use of a key principle of many real-world graphs, i.e., homophily, or the principle that ``like attracts like,'' as the guidance to effectively search various self-supervised pretext tasks. We provide theoretical understanding and empirical evidence to justify the flexibility of homophily in this search task. Then we propose the AutoSSL framework which can automatically search over combinations of various self-supervised tasks. By evaluating the framework on 7 real-world datasets, our experimental results show that AutoSSL can significantly boost the performance on downstream tasks including node clustering and node classification compared with training under individual tasks.

****

Creating Training Sets via Weak Indirect Supervision

Jieyu Zhang,Bohan Wang,Xiangchen Song,Yujing Wang,Yaming Yang,Jing Bai,Alexander Ratner

Creating labeled training sets has become one of the major roadblocks in machine learning. To address this, recent Weak Supervision (WS) frameworks synthesize training labels from multiple potentially noisy supervision sources. However, existing frameworks are restricted to supervision sources that share the same output space as the target task. To extend the scope of usable sources, we formulate Weak Indirect Supervision (WIS), a new research problem for automatically synthesizing training labels based on indirect supervision sources that have different output label spaces. To overcome the challenge of mismatched output spaces, we develop a probabilistic modeling approach, PLRM, which uses user-provided label relations to model and leverage indirect supervision sources. Moreover, we provide a theoretically-principled test of the distinguishability of PLRM for unseen labels, along with an generalization bound. On both image and text classification tasks as well as an industrial advertising application, we demonstrate the advantages of PLRM by outperforming baselines by a margin of 2%-9%.

****

PERSONALIZED LAB TEST RESPONSE PREDICTION WITH KNOWLEDGE AUGMENTATION

Suman Bhoi,Mong-Li Lee,Wynne Hsu,Hao Sen Andrew Fang,Ngiap Chuan Tan

Personalized medical systems are rapidly gaining traction as opposed to “one size fits all” systems. The ability to predict patients’ lab test responses and provide justification for the predictions would serve as an important decision support tool and help clinicians tailor treatment regimes for patients. This requires one to model the complex interactions among different medications, diseases, and lab tests. We also need to learn a strong patient representation, capturing both the sequential information accumulated over the visits and information from other similar patients. Further, we model the drug-lab interactions and diagnosis-lab interactions in the form of graphs and design a knowledge-augmented approach to predict patients’ response to a target lab result. We also take into consideration patients' past lab responses to personalize the prediction. Experiments on the benchmark MIMIC-III and a real-world outpatient dataset demonstrate the effectiveness of the proposed solution in reducing prediction errors by a significant margin. Case studies show that the identified top factors for influencing the predicted lab results are consistent with the clinicians' understanding.

****

Iterative Memory Network for Long Sequential User Behavior Modeling in Recommender Systems

Qianying Lin,Wen-Ji Zhou,Yanshi Wang,Qing Da,Qing-Guo Chen,Bing Wang

Sequential user behavior modeling is a key feature in modern recommender systems, seeking to capture users' interest based on their past activities. There are two usual approaches to sequential modeling : Recurrent Neural Networks (RNNs) and the attention mechanism. As the user behavior sequence gets longer, the usual approaches encounter problems. RNN-based methods incur the problem of fast forgetting, making it difficult to model the user's interests long time ago. The self-attention mechanism and its variations such as the transformer structure have the unfortunate property of a quadratic cost with respect to the input length, which makes it difficult to deal with long inputs. The target attention mechanism, despite having only $O(L)$ memory and time complexity, cannot model intra-sequence dependencies. In this paper, we propose Iterative Memory Network (IMN), an end-to-end differentiable framework for long sequential user behavior modeling. In our model, the target item acts as a memory trigger, continuously eliciting relevant information from the long sequence to represent the user's memory on the particular target item. In the Iterative Memory Update module, the model walks over the long sequence multiple iterations and keeps a memory vector to memorize the content walked over. Within each iteration, the sequence interacts with both the target item and the current memory for both target-sequence relation modeling and intra-sequence relation modeling. The memory is updated after each iteration. The framework incurs only $O(L)$ memory and time complexity while reduces the maximum length of network signal travelling paths to $O(1)$, which is achieved by the self-attention mechanism with $O(L^2)$ complexity. Various designs of efficient self-attention mechanisms are at best $O(LlogL)$. Extensive empirical studies show that our method outperforms various state-of-the-art sequential modeling methods on both public and industrial datasets for long sequential user behavior modeling.

****

Towards Demystifying Representation Learning with Non-contrastive Self-supervision

Xiang Wang,Xinlei Chen,Simon Shaolei Du,Yuandong Tian

Non-contrastive methods of self-supervised learning (such as BYOL and SimSiam) learn representations by minimizing the distance between two views of the same image. These approaches have achieved remarkable performance in practice, but it is not well understood 1) why these methods do not collapse to the trivial solutions and 2) how the representation is learned. Tian et al made an initial attempt on the first question and proposed DirectPred that sets the predictor directly. In our work, we analyze a generalized version of DirectPred, called DirectSet(α). We show that in a simple linear network, DirectSet(α) provably learns a desirable projection matrix and also reduces the sample complexity on downstream tasks. Our analysis suggests that weight decay acts as an implicit threshold that discard the features with high variance under augmentation, and keep the features with low variance. Inspired by our theory, we simplify DirectPred by removing the expensive eigen-decomposition step. On CIFAR-10, CIFAR-100, STL-10 and ImageNet, DirectCopy, our simpler and more computationally efficient algorithm rivals or even outperforms DirectPred.

****

Omni-Dimensional Dynamic Convolution

Chao Li,Aojun Zhou,Anbang Yao

Learning a single static convolutional kernel in each convolutional layer is the common training paradigm of modern Convolutional Neural Networks (CNNs). Instead, recent research in dynamic convolution shows that learning a linear combination of n convolutional kernels weighted with their input-dependent attentions can significantly improve the accuracy of light-weight CNNs, while maintaining efficient inference. However, we observe that existing works endow convolutional kernels with the dynamic property through one dimension (regarding the convolutiona

l kernel number) of the kernel space, but the other three dimensions (regarding the spatial size, the input channel number and the output channel number for each convolutional kernel) are overlooked. Inspired by this, we present Omni-dimensional Dynamic Convolution (ODConv), a more generalized yet elegant dynamic convolution design, to advance this line of research. ODConv leverages a novel multi-dimensional attention mechanism with a parallel strategy to learn complementary attentions for convolutional kernels along all four dimensions of the kernel space at any convolutional layer. As a drop-in replacement of regular convolutions, ODConv can be plugged into many CNN architectures. Extensive experiments on the ImageNet and MS-COCO datasets show that ODConv brings solid accuracy boosts for various prevailing CNN backbones including both light-weight and large ones, e.g., 3.77%~5.71%|1.86%~3.72% absolute top-1 improvements to MobivleNetV2|ResNet family on the ImageNet dataset. Intriguingly, thanks to its improved feature learning ability, ODConv with even one single kernel can compete with or outperform existing dynamic convolution counterparts with multiple kernels, substantially reducing extra parameters. Furthermore, ODConv is also superior to other attention modules for modulating the output features or the convolutional weights. Code and models will be available at https://github.com/OSVAI/ODConv.
****

Do Not Escape From the Manifold: Discovering the Local Coordinates on the Latent Space of GANs

Jaewoong Choi,Junho Lee,Changyeon Yoon,Jung Ho Park,Geonho Hwang,Myungjoo Kang

The discovery of the disentanglement properties of the latent space in GANs motivated a lot of research to find the semantically meaningful directions on it. In this paper, we suggest that the disentanglement property is closely related to the geometry of the latent space. In this regard, we propose an unsupervised method for finding the semantic-factorizing directions on the intermediate latent space of GANs based on the local geometry. Intuitively, our proposed method, called $\textit{Local Basis}$, finds the principal variation of the latent space in the neighborhood of the base latent variable. Experimental results show that the local principal variation corresponds to the semantic factorization and traversing along it provides strong robustness to image traversal. Moreover, we suggest an explanation for the limited success in finding the global traversal directions in the latent space, especially $\mathcal{W}$-space of StyleGAN2. We show that $\mathcal{W}$-space is warped globally by comparing the local geometry, discovered from Local Basis, through the metric on Grassmannian Manifold. The global warpage implies that the latent space is not well-aligned globally and therefore the global traversal directions are bound to show limited success on it.
****

A Systematic Evaluation of Domain Adaptation Algorithms On Time Series Data

Mohamed Ragab,Emadeldeen Eldele,Wee Ling Tan,Chuan-Sheng Foo,Zhenghua Chen,Min Wu,Chee Kwoh,Xiaoli Li

Unsupervised domain adaptation methods aim to generalize well on unlabeled test data that may have a different (shifted) distribution from the training data. Such methods are typically developed on image data, and their application to time series data is less explored. Existing works on time series domain adaptation suffer from inconsistencies in evaluation schemes, datasets, and base neural network architectures. Moreover, labeled target data are usually employed for model selection, which violates the fundamental assumption of unsupervised domain adaptation. To address these issues, we propose AdaTime, a standard framework to systematically and fairly evaluate different domain adaptation methods on time series data. Specifically, we standardize the base neural network architectures and benchmarking datasets, while also exploring more realistic model selection approaches that can work with no labeled data or few labeled samples. Our evaluation includes adaptations of state-of-the-art visual domain adaptation methods to time series data in addition to recent methods specifically developed for time series data. We conduct extensive experiments to evaluate 10 state-of-the-art methods on 3 representative datasets spanning 15 cross-domain scenarios. Our results suggest that with careful selection of hyper-parameters, visual domain adaptation methods are competitive with methods proposed for time series domain adaptation.

In addition, we find that model selection plays a key role and different selection strategies can significantly affect performance. Our work unveils practical insights for applying domain adaptation methods on time series data and builds a solid foundation for future works in the field.

****

Learning the Representation of Behavior Styles with Imitation Learning

Xiao Liu,Meng Wang,Zhaorong Wang,Yingfeng Chen,Yujing Hu,Changjie Fan,Chongjie Zhang

Imitation learning is one of the methods for reproducing expert demonstrations adaptively by learning a mapping between observations and actions. However, behavior styles such as motion trajectory and driving habit depend largely on the dataset of human maneuvers, and settle down to an average behavior style in most imitation learning algorithms. In this study, we propose a method named style behavior cloning (Style BC), which can not only infer the latent representation of behavior styles automatically, but also imitate different style policies from expert demonstrations. Our method is inspired by the word2vec algorithm and we construct a behavior-style to action mapping which is similar to the word-embedding to context mapping in word2vec. Empirical results on popular benchmark environments show that Style BC outperforms standard behavior cloning in prediction accuracy and expected reward significantly. Furthermore, compared with various baselines, our policy influenced by its assigned style embedding can better reproduce the expert behavior styles, especially in the complex environments or the number of the behavior styles is large.

****

GradSign: Model Performance Inference with Theoretical Insights

Zhihao Zhang,Zhihao Jia

A key challenge in neural architecture search (NAS) is quickly inferring the predictive performance of a broad spectrum of networks to discover statistically accurate and computationally efficient ones. We refer to this task as model performance inference (MPI). The current practice for efficient MPI is gradient-based methods that leverage the gradients of a network at initialization to infer its performance. However, existing gradient-based methods rely only on heuristic metrics and lack the necessary theoretical foundations to consolidate their designs. We propose GradSign, an accurate, simple, and flexible metric for model performance inference with theoretical insights. The key idea behind GradSign is a quantity Ψ to analyze the sample-wise optimization landscape of different networks. Theoretically, we show that Ψ is an upper bound for both the training and true population losses of a neural network under reasonable assumptions. However, it is computationally prohibitive to directly calculate Ψ for modern neural networks. To
address this challenge, we design GradSign, an accurate and simple approximation of Ψ using the gradients of a network evaluated at a random initialization state. Evaluation on seven NAS benchmarks across three training datasets shows that GradSign generalizes well to real-world networks and consistently outperforms state-of-the-art gradient-based methods for MPI evaluated by Spearman’s ρ and Kendall’s Tau. Additionally, we integrate GradSign into four existing NAS algorithms and show that the GradSign-assisted NAS algorithms outperform their vanilla counterparts by improving the accuracies of best-discovered networks by up to 0.3%, 1.1%, and 1.0% on three real-world tasks. Code is available at https://github.com/JackFram/GradSign

****

SiT: Simulation Transformer for Particle-based Physics Simulation

Yidi Shao,Chen Change Loy,Bo Dai

Most existing particle-based simulators adopt graph convolutional networks (GCNs) to model the underlying physics of particles.

However, they force particles to interact with all neighbors without selection, and they fall short in capturing material semantics for different particles, leading to unsatisfactory performance, especially in generalization.

This paper proposes Simulation Transformer (SiT) to simulate particle dynamics with more careful modeling of particle states, interactions, and their intrinsic

properties.
Specifically, besides the particle tokens, SiT generates interaction tokens and selectively focuses on essential interactions by allowing both tokens to attend to each other.
In addition, SiT learns material-aware representations by learnable abstract tokens, which will participate in the attention mechanism and boost the generalization capability further.
We evaluate our model on diverse environments, including fluid, rigid, and deformable objects, which cover systems of different complexity and materials.
Without bells and whistles, SiT shows strong abilities to simulate particles of different materials and achieves superior performance and generalization across these environments with fewer parameters than existing methods. Codes and models will be released.
****

Single-Cell Capsule Attention : an interpretable method of cell type classification for single-cell RNA-sequencing data
Tianxu Wang,Xiuli Ma
Single-cell RNA-sequencing technique can obtain genes’ expression level of every cell. Cell type classification (also known as cell type annotation) on single-cell RNA-seq data helps to explore cellular heterogeneity and diversity. Previous methods for cell type classification are either based on statistical hypotheses of gene expression or deep neural networks. However, the hypotheses may not reflect the true expression level. Deep neural networks lack interpretation for the result. Here we present an interpretable neural-network based method single-cell capsule attention(scCA) which assigns cells to different cell types based on their different feature patterns. In our model, we first generate capsules which extract different features of the cells. Then we obtain compound features which combine a set of features’ information through a LSTM model. In the end, we train attention weights and apply them to the compound features. scCA provides a strong interpretation for cell type classification result. Cells from the same cell type share a similar pattern of capsules’ relationship and similar distribution of attention weights for compound features. Compared with previous methods for cell type classification on nine datasets, scCA shows high accuracy on all datasets with robustness and reliable interpretation.
****

Learning Homophilic Incentives in Sequential Social Dilemmas
Heng Dong,Tonghan Wang,Jiayuan Liu,Chi Han,Chongjie Zhang
Promoting cooperation among self-interested agents is a long-standing and interdisciplinary problem, but receives less attention in multi-agent reinforcement learning (MARL). Game-theoretical studies reveal that altruistic incentives are critical to the emergence of cooperation but their analyses are limited to non-sequential social dilemmas. Recent works using deep MARL also show that learning to incentivize other agents has the potential to promote cooperation in more realistic sequential social dilemmas (SSDs). However, we find that, with these incentivizing mechanisms, the team cooperation level does not converge and regularly oscillates between cooperation and defection during learning. We show that a second-order social dilemma resulting from the incentive mechanisms is the main reason for such fragile cooperation. We analyze the dynamics of second-order social dilemmas and find that a typical tendency of humans, called homophily, provides a promising solution. We propose a novel learning framework to encourage homophilic incentives and show that it achieves stable cooperation in both SSDs of public goods and tragedy of the commons.
****

Variational Inference via Resolution of Singularities
Susan Wei
Predicated on the premise that neural networks are best viewed as singular statistical models, we set out to propose a new variational approximation for Bayesian neural networks. The approximation relies on a central result from singular learning theory according to which the posterior distribution over the parameters of a singular model, following an algebraic-geometrical transformation known as

a desingularization map, is asymptotically a mixture of standard forms. From here we proceed to demonstrate that a generalized gamma mean-field variational family, following desingularization, can recover the leading order term of the model evidence. Affine coupling layers are employed to learn the unknown desingularization map, effectively rendering the proposed methodology a normalizing flow with the generalized gamma as the source distribution.

****

A multi-domain splitting framework for time-varying graph structure

Zehua Yu,Xianwei Zheng,Zhulun Yang,Xutao Li

The Graph Signal Processing (GSP) methods are widely used to solve structured data analysis problems, assuming that the data structure is fixed. In the recent GSP community, anomaly detection on datasets with the time-varying structure is an open challenge. To address the anomaly detection problem for datasets with a spatial-temporal structure, in this work, we propose a novel graph multi-domain splitting framework, called GMDS, by integrating the time, vertex, and frequency features to locate the anomalies. Firstly, by introducing the discrete wavelet transform into vertex function, we design a splitting approach for separating the graph sequences into several sub-sequences adaptively. Then, we specifically design an adjacency function in the vertex domain to generate the adjacency matrix adaptively. At last, by utilizing the learned graphs to the spectral graph wavelet transform, we design a module to extract vertices features in the frequency domain. To validate the effectiveness of our framework, we apply GMDS in the anomaly detection of actual traffic flow and urban datasets and compare its performances with acknowledged baselines. The experimental results show that our proposed framework outperforms all the baselines, which distinctly demonstrate the validity of GMDS.

****

EViT: Expediting Vision Transformers via Token Reorganizations

Youwei Liang,Chongjian GE,Zhan Tong,Yibing Song,Jue Wang,Pengtao Xie

Vision Transformers (ViTs) take all the image patches as tokens and construct multi-head self-attention (MHSA) among them. Complete leverage of these image tokens brings redundant computations since not all the tokens are attentive in MHSA. Examples include that tokens containing semantically meaningless or distractive image backgrounds do not positively contribute to the ViT predictions. In this work, we propose to reorganize image tokens during the feed-forward process of ViT models, which is integrated into ViT during training. For each forward inference, we identify the attentive image tokens between MHSA and FFN (i.e., feed-forward network) modules, which is guided by the corresponding class token attention. Then, we reorganize image tokens by preserving attentive image tokens and fusing inattentive ones to expedite subsequent MHSA and FFN computations. To this end, our method EViT improves ViTs from two perspectives. First, under the same amount of input image tokens, our method reduces MHSA and FFN computation for efficient inference. For instance, the inference speed of DeiT-S is increased by 50% while its recognition accuracy is decreased by only 0.3% for ImageNet classification. Second, by maintaining the same computational cost, our method empowers ViTs to take more image tokens as input for recognition accuracy improvement, where the image tokens are from higher resolution images. An example is that we improve the recognition accuracy of DeiT-S by 1% for ImageNet classification at the same computational cost of a vanilla DeiT-S. Meanwhile, our method does not introduce more parameters to ViTs. Experiments on the standard benchmarks show the effectiveness of our method. The code is available at https://github.com/youweiliang/evit

****

Lottery Ticket Structured Node Pruning for Tabular Datasets

Ryan Bluteau,Robin Gras,Mitchel Paulin,Zachary Innes

In this paper we presented two pruning approaches on tabular neural networks based on the lottery ticket hypothesis that went beyond masking nodes by resizing the models accordingly. We showed top performing models in 6 of 8 datasets tested in terms of F1/RMSE. We also showed in 6 of 8 datasets a total reduction of over 85% of nodes and many over 98% reduced with minimal affect to accuracy. In one dataset the model reached a total size of one node per layer while still improving RMSE compared to the larger model used for pruning. We presented results for two approaches, iterative pruning using two styles, and oneshot pruning. Iterative pruning gradually reduces nodes in each layers based on norm pruning until we reach the smallest state, while oneshot will prune the model directly to the smallest state. We showed that the iterative approach will obtain the best result more consistently than oneshot.

****

Hierarchical Modular Framework for Long Horizon Instruction Following

Suvaansh Bhambri,Byeonghwi Kim,Roozbeh Mottaghi,Jonghyun Choi

Robotic agents performing domestic chores using natural language directives re-quire to learn the complex task of navigating an environment and interacting with objects in it. To address such composite tasks, we propose a hierarchical modular approach to learn agents that navigate and manipulate objects in a divide-and-conquer manner for the diverse nature of the entailing tasks. Specifically, our policy operates at three levels of hierarchy. We first infer a sequence of subgoals to be executed based on language instructions by a high-level policy composition controller (PCC). We then discriminatively control the agent’s navigation by a master policy by alternating between navigation policy and various independent interaction policies. Finally, we infer manipulation actions with the corresponding object masks using the appropriate interaction policy. Our hierarchical agent, named HACR (Hierarchical Approach for Compositional Reasoning), generates a human interpretable and short sequence of sub-objectives, leading to efficient interaction with an environment, and achieves the state-of-the-art performance on the challenging ALFRED benchmark.

****

You are AllSet: A Multiset Function Framework for Hypergraph Neural Networks

Eli Chien,Chao Pan,Jianhao Peng,Olgica Milenkovic

Hypergraphs are used to model higher-order interactions amongst agents and there exist many practically relevant instances of hypergraph datasets. To enable the efficient processing of hypergraph data, several hypergraph neural network platforms have been proposed for learning hypergraph properties and structure, with a special focus on node classification tasks. However, almost all existing methods use heuristic propagation rules and offer suboptimal performance on benchmarking datasets. We propose AllSet, a new hypergraph neural network paradigm that represents a highly general framework for (hyper)graph neural networks and for the first time implements hypergraph neural network layers as compositions of two multiset functions that can be efficiently learned for each task and each dataset. The proposed AllSet framework also for the first time integrates Deep Sets and Set Transformers with hypergraph neural networks for the purpose of learning multiset functions and therefore allows for significant modeling flexibility and high expressive power. To evaluate the performance of AllSet, we conduct the most extensive experiments to date involving ten known benchmarking datasets and three newly curated datasets that represent significant challenges for hypergraph node classification. The results demonstrate that our method has the unique ability to either match or outperform all other hypergraph neural networks across the tested datasets: As an example, the performance improvements over existing methods and a new method based on heterogeneous graph neural networks are close to 4% on the Yelp and Zoo datasets, and 3% on the Walmart dataset.

****

Open Set Domain Adaptation with Zero-shot Learning on Graph

Xinyue Zhang,Xu Yang,Zhi-yong Liu

Open set domain adaptation focuses on transferring the information from a richly labeled domain called \emph{source domain} to a scarcely labeled domain called \emph{target domain} while classifying the unseen target samples as one \emph{unknown} class in an unsupervised way. Compared with the close set domain adaptation, where the source domain and the target domain share the same class space, the classification of the unknown class makes it easy to adapt to the realistic environment. Particularly, after the recognition of the unknown samples, the robot can either ask for manually labeling or further develop the classification abil ity of the unknown classes based on pre-stored knowledge. Inspired by this idea, in this paper we propose a model for open set domain adaptation with zero-shot learning on the unknown classes. We utilize adversarial learning to align the two domains while rejecting the unknown classes. Then the knowledge graph is introduced to generate the classifiers for the unknown classes with the employment of the graph convolution network (GCN). Thus the classification ability of the source domain is transferred to the target domain and the model can distinguish the unknown classes with prior knowledge. We evaluate our model on digits datasets and the result shows superior performance.

****

FEDERATED LEARNING FRAMEWORK BASED ON TRIMMED MEAN AGGREGATION RULES

Wang Tian Xiang,Meiyue Shao,Yanwei Fu,Riheng Jia,Feilong Lin,Zhonglong Zheng

This paper studies the problem of information security in the distributed learning framework. In particular, we consider the clients will always be attacked by Byzantine nodes and poisoning in the federated learning. Typically, aggregation rules are utilized to protect the model from the attacks in federated learning. The classical aggregation methods are Krum(·) and Mean(·), which however, are not capable enough to deal with Byzantine attacks in which general deviations and multiple clients are attacked at the same time. We propose new aggregation rules, Tmean(·), to the federated learning algorithm, and propose a federated learning framework based on Byzantine-resilient aggregation algorithm. Our novel Tmean(·) rules are derived from Mean(·) by appropriately trimming some of the values before averaging them. Theoretically, we provide rigorous theoretical proof and understanding of Tmean(·). Extensive experiments validate the effectiveness of our approaches.

****

A Theoretical and Empirical Model of the Generalization Error under Time-Varying Learning Rate

Toru Makuuchi,YUSUKE Mukuta,Tatsuya Harada

Stochastic gradient descent is commonly employed as the most principled optimization algorithm for deep learning, and the dependence of the generalization error of neural networks on the given hyperparameters is crucial.

However, the case in which the batch size and learning rate vary with time has not yet been analyzed, nor the dependence of them on the generalization error as a functional form for both the constant and time-varying cases has been expressed.

In this study, we analyze the generalization bound for the time-varying case by applying PAC-Bayes and experimentally show that the theoretical functional form for the batch size and learning rate approximates the generalization error well for both cases.

We also experimentally show that hyperparameter optimization based on the proposed model outperforms the existing libraries.

****

Synchromesh: Reliable Code Generation from Pre-trained Language Models

Gabriel Poesia,Alex Polozov,Vu Le,Ashish Tiwari,Gustavo Soares,Christopher Meek,Sumit Gulwani

Large pre-trained language models have been used to generate code, providing a flexible interface for synthesizing programs from natural language specifications. However, they often violate syntactic and semantic rules of their output language, limiting their practical usability. In this paper, we propose Synchromesh: a framework for substantially improving the reliability of pre-trained models for code generation. Synchromesh comprises two components. First, it retrieves few-shot examples from a training bank using Target Similarity Tuning (TST), a novel method for semantic example selection. TST learns to recognize utterances that describe similar target programs despite of differences in surface natural language features. Then, Synchromesh feeds the examples to a pre-trained language model and samples programs using Constrained Semantic Decoding (CSD): a general framework for constraining the output to a set of valid programs in the target language. CSD leverages constraints on partial outputs to sample complete correct programs, and needs neither re-training nor fine-tuning of the language model. We evaluate our methods by synthesizing code from natural language descriptions using GPT-3 and Codex in three real-world languages: SQL queries, Vega-Lite visualizations and SMCalFlow programs. These domains showcase rich constraints that CSD is able to enforce, including syntax, scoping and typing rules. Across all languages, we observe complementary gains from CSD and TST in prediction accuracy and in effectively preventing parsing, type and run-time errors.

****

Learning curves for continual learning in neural networks: Self-knowledge transfer and forgetting

Ryo Karakida,Shotaro Akaho

Sequential training from task to task is becoming one of the major objects in deep learning applications such as continual learning and transfer learning. Nevertheless, it remains unclear under what conditions the trained model's performance improves or deteriorates. To deepen our understanding of sequential training, this study provides a theoretical analysis of generalization performance in a solvable case of continual learning. We consider neural networks in the neural tangent kernel (NTK) regime that continually learn target functions from task to task, and investigate the generalization by using an established statistical mechanical analysis of kernel ridge-less regression. We first show characteristic transitions from positive to negative transfer. More similar targets above a specific critical value can achieve positive knowledge transfer for the subsequent task while catastrophic forgetting occurs even with very similar targets. Next, we investigate a variant of continual learning which supposes the same target function in multiple tasks. Even for the same target, the trained model shows some transfer and forgetting depending on the sample size of each task. We can guarantee that the generalization error monotonically decreases from task to task for equal sample sizes while unbalanced sample sizes deteriorate the generalization. We respectively refer to these improvement and deterioration as self-knowledge transfer and forgetting, and empirically confirm them in realistic training of deep neural networks as well.

****

Energy-Based Learning for Cooperative Games, with Applications to Valuation Problems in Machine Learning

Yatao Bian,Yu Rong,Tingyang Xu,Jiaxiang Wu,Andreas Krause,Junzhou Huang

Valuation problems, such as feature interpretation, data valuation and model valuation for ensembles, become increasingly more important in many machine learning applications. Such problems are commonly solved by well-known game-theoretic criteria, such as Shapley value or Banzhaf value. In this work, we present a novel energy-based treatment for cooperative games, with a theoretical justification by the maximum entropy framework. Surprisingly, by conducting variational inference of the energy-based model, we recover various game-theoretic valuation criteria through conducting one-step fixed point iteration for maximizing the mean-field ELBO objective. This observation also verifies the rationality of existing criteria, as they are all attempting to decouple the correlations among the players through the mean-field approach. By running fixed point iteration for multiple steps, we achieve a trajectory of the valuations, among which we define the valuation with the best conceivable decoupling error as the Variational Index. We prove that under uniform initializations, these variational valuations all satisfy a set of game-theoretic axioms. We experimentally demonstrate that the proposed Variational Index enjoys lower decoupling error and better valuation performance on certain synthetic and real-world valuation problems.

****

Pessimistic Model-based Offline Reinforcement Learning under Partial Coverage

Masatoshi Uehara,Wen Sun

We study model-based offline Reinforcement Learning with general function approximation without a full coverage assumption on the offline data distribution. We present an algorithm named Constrained Pessimistic Policy Optimization (CPPO) which leverages a general function class and uses a constraint over the models to encode pessimism. Under the assumption that the ground truth model belongs to our function class (i.e., realizability in the function class), CPPO has a PAC gua rantee with offline data only providing partial coverage, i.e., it can learn a policy that competes against any policy covered by the offline data. We then demonstrate that this algorithmic framework can be applied to many specialized Markov Decision Processes where the additional structural assumptions can further refine the concept of partial coverage. Two notable examples are: (1) low- rank MDP with representation learning where the partial coverage condition is defined using a relative condition number measured by the unknown ground truth feature representation; (2) factored MDP where the partial coverage condition is defined using density-ratio based concentrability coefficients associated with individual factors.

****

RL-DARTS: Differentiable Architecture Search for Reinforcement Learning

Yingjie Miao,Xingyou Song,Daiyi Peng,Summer Yue,John D Co-Reyes,Eugene Brevdo,Aleksandra Faust

Recently, Differentiable Architecture Search (DARTS) has become one of the most popular Neural Architecture Search (NAS) methods successfully applied in supervised learning (SL). However, its applications in other domains, in particular for reinforcement learning (RL), has seldom been studied. This is due in part to RL possessing a significantly different optimization paradigm than SL, especially with regards to the notion of replay data, which is continually generated via inference in RL. In this paper, we introduce RL-DARTS, one of the first applications of end-to-end DARTS in RL to search for convolutional cells, applied to the challenging, infinitely procedurally generated Procgen benchmark. We demonstrate that the benefits of DARTS become amplified when applied to RL, namely search efficiency in terms of time and compute, as well as simplicity in integration with complex preexisting RL code via simply replacing the image encoder with a DARTS supernet, compatible with both off-policy and on-policy RL algorithms. At the same time however, we provide one of the first extensive studies of DARTS outside of the standard fixed dataset setting in SL via RL-DARTS. We show that throughout training, the supernet gradually learns better cells, leading to alternative architectures which can be highly competitive against manually designed policies, but also verify previous design choices for RL policies.

****

A new look at fairness in stochastic multi-armed bandit problems

Guanhua Fang,Ping Li,Gennady Samorodnitsky

We study an important variant of the stochastic multi-armed bandit (MAB) problem, which takes fairness into consideration. Instead of directly maximizing cumulative expected reward, we need to balance between the total reward and fairness level. In this paper, we present a new insight in MAB with fairness and formulate the problem in the penalization framework, where rigorous penalized regret can be well defined and more sophisticated regret analysis is possible. Under such a framework, we propose a hard-threshold UCB-like algorithm, which enjoys many merits including asymptotic fairness, nearly optimal regret, better tradeoff between reward and fairness. Both gap-dependent and gap-independent upper bounds have been established. Lower bounds are also given to illustrate the tightness of our theoretical analysis. Numerous experimental results corroborate the theory and show the superiority of our method over other existing methods.

****

Evaluating Robustness of Cooperative MARL

Nhan Pham,Lam M. Nguyen,Jie Chen,Thanh Lam Hoang,Subhro Das,Tsui-Wei Weng

In recent years, a proliferation of methods were developed for multi-agent reinforcement learning (MARL). In this paper, we focus on evaluating the robustness of MARL agents in continuous control tasks. In particular, we propose the first model-based approach to perform adversarial attacks for cooperative MARL. We design effective attacks to degrade the MARL agent's performance by adversarially perturbing the states of agent(s) and solving an optimization problem. In addition, we also developed several strategies to select the most vulnerable agents that help to further decrease the team reward of MARL. Extensive numerical experiments on multi-agent Mujoco tasks verify the effectiveness of our proposed approach.

****

New Perspective on the Global Convergence of Finite-Sum Optimization

Lam M. Nguyen,Trang H. Tran,Marten van Dijk

Deep neural networks (DNNs) have shown great success in many machine learning tasks. Their training is challenging since the loss surface of the network architecture is generally non-convex, or even non-smooth. How and under what assumptions is guaranteed convergence to a \textit{global} minimum possible? We propose a reformulation of the minimization problem allowing for a new recursive algorithmic framework. By using bounded style assumptions, we prove convergence to an ε-(global) minimum using $\mathcal{\tilde{O}}(1/\varepsilon^2)$ gradient computations. Our theoretical foundation motivates further study, implementation, and optimization of the new algorithmic framework and further investigation of its non-standard bounded style assumptions. This new direction broadens our understanding of why and under what circumstances training of a DNN converges to a global minimum.

****

Cold Brew: Distilling Graph Node Representations with Incomplete or Missing Neighborhoods

Wenqing Zheng,Edward W Huang,Nikhil Rao,Sumeet Katariya,Zhangyang Wang,Karthik Subbian

Graph Neural Networks (GNNs) have achieved state-of-the-art performance in node classification, regression, and recommendation tasks. GNNs work well when rich and high-quality connections are available. However, their effectiveness is often jeopardized in many real-world graphs in which node degrees have power-law distributions. The extreme case of this situation, where a node may have no neighbors, is called Strict Cold Start (SCS). SCS forces the prediction to rely completely on the node's own features. We propose Cold Brew, a teacher-student distillation approach to address the SCS and noisy-neighbor challenges for GNNs. We also introduce feature contribution ratio (FCR), a metric to quantify the behavior of inductive GNNs to solve SCS. We experimentally show that FCR disentangles the contributions of different graph data components and helps select the best architecture for SCS generalization. We further demonstrate the superior performance of Cold Brew on several public benchmark and proprietary e-commerce datasets, where many nodes have either very few or noisy connections. Our source code is available at https://github.com/amazon-research/gnn-tail-generalization.

****

NASI: Label- and Data-agnostic Neural Architecture Search at Initialization

Yao Shu,Shaofeng Cai,Zhongxiang Dai,Beng Chin Ooi,Bryan Kian Hsiang Low

Recent years have witnessed a surging interest in Neural Architecture Search (NAS). Various algorithms have been proposed to improve the search efficiency and effectiveness of NAS, i.e., to reduce the search cost and improve the generalization performance of the selected architectures, respectively. However, the search efficiency of these algorithms is severely limited by the need for model training during the search process. To overcome this limitation, we propose a novel NAS algorithm called NAS at Initialization (NASI) that exploits the capability of a Neural Tangent Kernel in being able to characterize the performance of candidate architectures at initialization, hence allowing model training to be completely avoided to boost the search efficiency. Besides the improved search efficiency, NASI also achieves competitive search effectiveness on various datasets like CIFAR-10/100 and ImageNet. Further, NASI is shown to be label- and data-agnostic under mild conditions, which guarantees the transferability of architectures selected by our NASI over different datasets.

****

Structured Stochastic Gradient MCMC

Antonios Alexos,Alex James Boyd,Stephan Mandt

Stochastic gradient Markov Chain Monte Carlo (SGMCMC) is considered the gold standard for Bayesian inference in large-scale models, such as Bayesian neural networks. Since practitioners face speed versus accuracy tradeoffs in these models, variational inference (VI) is often the preferable option. Unfortunately, VI makes strong assumptions on both the factorization and functional form of the poste rior. In this work, we propose a new non-parametric variational approximation that makes no assumptions about the approximate posterior's functional form and allows practitioners to specify the exact dependencies the algorithm should respect or break. The approach relies on a new Langevin-type algorithm that operates on a modified energy function, where parts of the latent variables are averaged over samples from earlier iterations of the Markov chain. This way, statistical dependencies can be broken in a controlled way, allowing the chain to mix faster. This scheme can be further modified in a ``dropout'' manner, leading to even more scalability. By implementing the scheme on a ResNet-20 architecture, we obtain better predictive likelihoods and faster mixing time than full SGMCMC.

****

f-Mutual Information Contrastive Learning
Guojun Zhang,Yiwei Lu,Sun Sun,Hongyu Guo,Yaoliang Yu

Self-supervised contrastive learning is an emerging field due to its power in providing good data representations. Such learning paradigm widely adopts the InfoNCE loss, which is closely connected with maximizing the mutual information. In this work, we propose the f-Mutual Information Contrastive Learning framework (f-MICL) , which directly maximizes the f-divergence-based generalization of mutual information. We theoretically prove that, under mild assumptions, our f-MICL naturally attains the alignment for positive pairs and the uniformity for data representations, the two main factors for the success of contrastive learning. We further provide theoretical guidance on designing the similarity function and choosing the effective f-divergences for f-MICL. Using several benchmark tasks from both vision and natural text, we empirically verify that our novel method outperforms or performs on par with state-of-the-art strategies.

****

How to Train Your MAML to Excel in Few-Shot Classification
Han-Jia Ye,Wei-Lun Chao

Model-agnostic meta-learning (MAML) is arguably one of the most popular meta-learning algorithms nowadays.
Nevertheless, its performance on few-shot classification is far behind many recent algorithms dedicated to the problem. In this paper, we point out several key facets of how to train MAML to excel in few-shot classification. First, we find that MAML needs a large number of gradient steps in its inner loop update, which contradicts its common usage in few-shot classification. Second, we find that MAML is sensitive to the class label assignments during meta-testing. Concretely, MAML meta-trains the initialization of an N-way classifier. These N ways, during meta-testing, then have "$N!$" different permutations to be paired with a few-shot task of N novel classes. We find that these permutations lead to a huge variance of accuracy, making MAML unstable in few-shot classification. Third, we investigate several approaches to make MAML permutation-invariant, among which meta-training a single vector to initialize all the N weight vectors in the classification head performs the best. On benchmark datasets like MiniImageNet and TieredImageNet, our approach, which we name UNICORN-MAML, performs on a par with or even outperforms many recent few-shot classification algorithms, without sacrificing MAML's simplicity.

****

Communication-Efficient Actor-Critic Methods for Homogeneous Markov Games
Dingyang Chen,Yile Li,Qi Zhang

Recent success in cooperative multi-agent reinforcement learning (MARL) relies on centralized training and policy sharing. Centralized training eliminates the issue of non-stationarity MARL yet induces large communication costs, and policy sharing is empirically crucial to efficient learning in certain tasks yet lacks theoretical justification. In this paper, we formally characterize a subclass of cooperative Markov games where agents exhibit a certain form of homogeneity such that policy sharing provably incurs no suboptimality. This enables us to develop the first consensus-based decentralized actor-critic method where the consensus update is applied to both the actors and the critics while ensuring convergence. We also develop practical algorithms based on our decentralized actor-critic method to reduce the communication cost during training, while still yielding p olicies comparable with centralized training.

****

FedDiscrete: A Secure Federated Learning Algorithm Against Weight Poisoning
Yutong Dai,Xingjun Ma,Lichao Sun
Federated learning (FL) is a privacy-aware collaborative learning paradigm that allows multiple parties to jointly train a machine learning model without sharing their private data. However, recent studies have shown that FL is vulnerable to weight poisoning attacks. In this paper, we propose a probabilistic discretization mechanism on the client side, which transforms the client's model weight into a vector that can only have two different values but still guarantees that the server obtains an unbiased estimation of the client's model weight. We theoretically analyze the utility, robustness, and convergence of our proposed discretization mechanism and empirically verify its superior robustness against various weight-based attacks under the cross-device FL setting.

****

Self-Supervised Representation Learning via Latent Graph Prediction
Yaochen Xie,Zhao Xu,Shuiwang Ji
Self-supervised learning (SSL) of graph neural networks is emerging as a promising way of leveraging unlabeled data. Currently, most methods are based on contrastive learning adapted from the image domain, which requires view generation and a sufficient number of negative samples. In contrast, existing predictive models do not require negative sampling, but lack theoretical guidance on the design of pretext training tasks. In this work, we propose the LaGraph, a theoretically grounded predictive SSL framework based on latent graph prediction. Learning objectives of LaGraph are derived as self-supervised upper bounds to objectives for predicting unobserved latent graphs. In addition to its improved performance, LaGraph provides explanations for recent successes of predictive models that include invariance-based objectives. We provide theoretical analysis comparing LaGraph to related methods in different domains. Our experimental results demonstrate the superiority of LaGraph in performance and the robustness to decreasing of training sample size on both graph-level and node-level tasks.

****

S$^3$ADNet: Sequential Anomaly Detection with Pessimistic Contrastive Learning
Quexuan Zhang,Yukio Ohsawa
Anomalies are commonly found in sequential data generated by real-world applications, such as cyberattacks in network traffic, human activity changes in wearable sensors. Thanks to the development of computing technology, many impressive results have been obtained from deep learning-based anomaly detection approaches in recent years. This paper proposes a simple neural network framework for detecting anomalies on sequential data, called $S$elf-$S$upervised $S$equential $A$nomaly $D$etection $N$etwork (S$^3$ADNet). S$^3$ADNet first extracts the representations from each data point by performing feature augmentation for contrastive learning; then captures the contextual information from the sequential data points for estimating anomaly probabilities by optimizing the context-adaptive objective. Here, we design a novel loss function based on a pessimistic policy, considering that only anomalies can affect the contextual relationships in sequences. Our proposed method outperformed other state-of-the-art approaches on the benchmark datasets by F1-score with a more straightforward architecture.

****

MobileViT: Light-weight, General-purpose, and Mobile-friendly Vision Transformer
Sachin Mehta,Mohammad Rastegari
Light-weight convolutional neural networks (CNNs) are the de-facto for mobile vision tasks. Their spatial inductive biases allow them to learn representations with fewer parameters across different vision tasks. However, these networks are spatially local. To learn global representations, self-attention-based vision trans-formers (ViTs) have been adopted. Unlike CNNs, ViTs are heavy-weight. In this paper, we ask the following question: is it possible to combine the strengths of CNNs and ViTs to build a light-weight and low latency network for mobile vision tasks? Towards this end, we introduce MobileViT, a light-weight and general-purpose vision transformer for mobile devices. MobileViT presents a different per

spective for the global processing of information with transformers, i.e., transformers as convolutions. Our results show that MobileViT significantly outperforms CNN- and ViT-based networks across different tasks and datasets. On the ImageNet-1k dataset, MobileViT achieves top-1 accuracy of 78.4% with about 6 million parameters, which is 3.2% and 6.2% more accurate than MobileNetv3 (CNN-based) and DeIT (ViT-based) for a similar number of parameters. On the MS-COCO object detection task, MobileViT is 5.7% more accurate than MobileNetv3 for a similar number of parameters.

Our source code is open-source and available at: https://github.com/apple/ml-cvnets

****

Spatial Graph Attention and Curiosity-driven Policy for Antiviral Drug Discovery

Yulun Wu,Nicholas Choma,Andrew Deru Chen,Mikaela Cashman,Erica Teixeira Prates,Veronica G Melesse Vergara,Manesh B Shah,Austin Clyde,Thomas Brettin,Wibe Albert de Jong,Neeraj Kumar,Martha S Head,Rick L. Stevens,Peter Nugent,Daniel A Jacobson,James B Brown

We developed Distilled Graph Attention Policy Network (DGAPN), a reinforcement learning model to generate novel graph-structured chemical representations that optimize user-defined objectives by efficiently navigating a physically constrained domain. The framework is examined on the task of generating molecules that are designed to bind, noncovalently, to functional sites of SARS-CoV-2 proteins. We present a spatial Graph Attention (sGAT) mechanism that leverages self-attention over both node and edge attributes as well as encoding the spatial structure --- this capability is of considerable interest in synthetic biology and drug discovery. An attentional policy network is introduced to learn the decision rules for a dynamic, fragment-based chemical environment, and state-of-the-art policy gradient techniques are employed to train the network with stability. Exploration is driven by the stochasticity of the action space design and the innovation reward bonuses learned and proposed by random network distillation. In experiments, our framework achieved outstanding results compared to state-of-the-art algorithms, while reducing the complexity of paths to chemical synthesis.

****

Surrogate NAS Benchmarks: Going Beyond the Limited Search Spaces of Tabular NAS Benchmarks

Arber Zela,Julien Niklas Siems,Lucas Zimmer,Jovita Lukasik,Margret Keuper,Frank Hutter

The most significant barrier to the advancement of Neural Architecture Search (NAS) is its demand for large computational resources, which hinders scientifically sound empirical evaluations of NAS methods. Tabular NAS benchmarks have alleviated this problem substantially, making it possible to properly evaluate NAS methods in seconds on commodity machines. However, an unintended consequence of tabular NAS benchmarks has been a focus on extremely small architectural search spaces since their construction relies on exhaustive evaluations of the space. This leads to unrealistic results that do not transfer to larger spaces. To overcome this fundamental limitation, we propose a methodology to create cheap NAS surrogate benchmarks for arbitrary search spaces. We exemplify this approach by creating surrogate NAS benchmarks on the existing tabular NAS-Bench-101 and on two widely used NAS search spaces with up to 10^{21} architectures (10^{13} times larger than any previous tabular NAS benchmark). We show that surrogate NAS benchmarks can model the true performance of architectures better than tabular benchmarks (at a small fraction of the cost), that they lead to faithful estimates of how well different NAS methods work on the original non-surrogate benchmark, and that they can generate new scientific insight. We open-source all our code and believe that surrogate NAS benchmarks are an indispensable tool to extend scientifically sound work on NAS to large and exciting search spaces.

****

Certified Robustness for Deep Equilibrium Models via Interval Bound Propagation

Colin Wei,J Zico Kolter

Deep equilibrium layers (DEQs) have demonstrated promising performance and are c ompetitive with standard explicit models on many benchmarks. However, little is known about certifying robustness for these models. Inspired by interval bound propagation (IBP), we propose the IBP-MonDEQ layer, a DEQ layer whose robustness can be verified by computing upper and lower interval bounds on the output. Our key insights are that these interval bounds can be obtained as the fixed-point solution to an IBP-inspired equilibrium equation, and furthermore, that this solution always exists and is unique when the layer obeys a certain parameterization. This fixed point can be interpreted as the result of applying IBP to an infinitely deep, weight-tied neural network, which may be of independent interest, as IBP bounds are typically unstable for deeper networks. Our empirical comparison reveals that models with IBP-MonDEQ layers can achieve comparable ℓ_{∞} certified robustness to similarly-sized fully explicit networks.

****

Crystal Diffusion Variational Autoencoder for Periodic Material Generation
Tian Xie,Xiang Fu,Octavian-Eugen Ganea,Regina Barzilay,Tommi S. Jaakkola

Generating the periodic structure of stable materials is a long-standing challenge for the material design community. This task is difficult because stable materials only exist in a low-dimensional subspace of all possible periodic arrangements of atoms: 1) the coordinates must lie in the local energy minimum defined by quantum mechanics, and 2) global stability also requires the structure to follow the complex, yet specific bonding preferences between different atom types. Existing methods fail to incorporate these factors and often lack proper invariances. We propose a Crystal Diffusion Variational Autoencoder (CDVAE) that captures the physical inductive bias of material stability. By learning from the data distribution of stable materials, the decoder generates materials in a diffusion process that moves atomic coordinates towards a lower energy state and updates atom types to satisfy bonding preferences between neighbors. Our model also explicitly encodes interactions across periodic boundaries and respects permutation, translation, rotation, and periodic invariances. We significantly outperform past methods in three tasks: 1) reconstructing the input structure, 2) generating valid, diverse, and realistic materials, and 3) generating materials that optimize a specific property. We also provide several standard datasets and evaluation metrics for the broader machine learning community.

****

Lottery Tickets can have Structural Sparsity
Tianlong Chen,Xuxi Chen,Xiaolong Ma,Yanzhi Wang,Zhangyang Wang

The lottery ticket hypothesis (LTH) has shown that dense models contain highly sparse subnetworks (i.e., $\textit{winning tickets}$) that can be trained in isolation to match full accuracy. Despite many exciting efforts being made, there is one "commonsense" seldomly challenged: a winning ticket is found by iterative magnitude pruning (IMP) and hence the resultant pruned subnetworks have only unstructured sparsity. That gap limits the appeal of winning tickets in practice, since the highly irregular sparse patterns are challenging to accelerate on hardware. Meanwhile, directly substituting structured pruning for unstructured pruning in IMP damages performance more severely and is usually unable to locate winning tickets.

In this paper, we demonstrate $\textbf{the first positive result}$ that a structurally sparse winning ticket can be effectively found in general. The core idea is to append ``post-processing techniques" after each round of (unstructured) IMP, to enforce the formation of structural sparsity. Specifically, we first ``re-fill" pruned elements back in some channels deemed to be important, and then ``re-group" non-zero elements to create flexible group-wise structural patterns. Both our identified channel- and group-wise structural subnetworks win the lottery, with substantial inference speedups readily supported by practical hardware. Extensive experiments, conducted on diverse datasets across multiple network backbones, consistently validate our proposal, showing that the hardware acceleration roadblock of LTH is now removed. Specifically, the structural winning tickets obtain up to $\{64.93\%, 64.84\%, 64.84\%\}$ running time savings at $\{36\%\sim 80\%, 74\%, 58\%\}$ sparsity on CIFAR, Tiny-ImageNet, ImageNet, while maintaini ng comparable accuracy. All the codes and pre-trained models will be publicly released.

****

Task Affinity with Maximum Bipartite Matching in Few-Shot Learning

Cat Phuoc Le,Juncheng Dong,Mohammadreza Soltani,Vahid Tarokh

We propose an asymmetric affinity score for representing the complexity of utilizing the knowledge of one task for learning another one. Our method is based on the maximum bipartite matching algorithm and utilizes the Fisher Information matrix. We provide theoretical analyses demonstrating that the proposed score is mathematically well-defined, and subsequently use the affinity score to propose a novel algorithm for the few-shot learning problem. In particular, using this score, we find relevant training data labels to the test data and leverage the discovered relevant data for episodically fine-tuning a few-shot model. Results on various few-shot benchmark datasets demonstrate the efficacy of the proposed approach by improving the classification accuracy over the state-of-the-art methods even when using smaller models.

****

Decentralized Cooperative Multi-Agent Reinforcement Learning with Exploration

Weichao Mao,Tamer Basar,Lin Yang,Kaiqing Zhang

Many real-world applications of multi-agent reinforcement learning (RL), such as multi-robot navigation and decentralized control of cyber-physical systems, involve the cooperation of agents as a team with aligned objectives. We study multi-agent RL in the most basic cooperative setting --- Markov teams --- a class of Markov games where the cooperating agents share a common reward. We propose an algorithm in which each agent independently runs stage-based V-learning (a Q-learning style algorithm) to efficiently explore the unknown environment, while using a stochastic gradient descent (SGD) subroutine for policy updates. We show that the agents can learn an ϵ-approximate Nash equilibrium policy in at most $\propto\widetilde{O}(1/\epsilon^4)$ episodes. Our results advocate the use of a novel \emph{stage-based} V-learning approach to create a stage-wise stationary environment. We also show that under certain smoothness assumptions of the team, our algorithm can achieve a nearly \emph{team-optimal} Nash equilibrium. Simulation results corroborate our theoretical findings. One key feature of our algorithm is being \emph{decentralized}, in the sense that each agent has access to only the state and its local actions, and is even \emph{oblivious} to the presence of the other agents. Neither communication among teammates nor coordination by a central controller is required during learning. Hence, our algorithm can readily generalize to an arbitrary number of agents, without suffering from the exponential dependence on the number of agents.

****

HODA: Protecting DNNs Against Model Extraction Attacks via Hardness of Samples

AmirMahdi Sadeghzadeh,Faezeh Dehghan,Amir Sobhanian,Rasool Jalili

Model Extraction attacks exploit the target model's prediction API to create a surrogate model in order to steal or reconnoiter the functionality of the target model in the black-box setting. Several recent studies have shown that a data-limited adversary who has no or limited access to the samples from the target model's training data distribution can use synthesis or semantically similar samples to conduct model extraction attacks. As the training process of DNN-based classifiers is done in several epochs, we can consider this process as a sequence of subclassifiers so that each subclassifier is created at the end of an epoch. We use the sequence of subclassifiers to calculate the hardness degree of samples. In this paper, we investigate the hardness degree of samples and demonstrate that the hardness degree histogram of a data-limited adversary's sample sequences is distinguishable from the hardness degree histogram of benign users' samples sequences, consisting of normal samples. Normal samples come from the target classifier's training data distribution. We propose Hardness-Oriented Detection Approach (HODA) to detect the sample sequences of model extraction attacks. The results demonstrate that HODA can detect the sample sequences of model extraction attacks with a high success rate by only watching 100 samples of them.

****

Invariance in Policy Optimisation and Partial Identifiability in Reward Learning
Joar Max Viktor Skalse,Matthew Farrugia-Roberts,Stuart Russell,Adam Gleave

It is challenging to design a reward function for complex, real-world tasks. Reward learning algorithms let one instead infer a reward function from data. However, multiple reward functions often explain the data equally well, even in the limit of infinite data. Prior work has focused on situations where the reward function is uniquely recoverable, by introducing additional assumptions or data sources. By contrast, we formally characterise this partial identifiability for popular data sources such as demonstrations and trajectory preferences. We analyse the impact of this ambiguity on downstream tasks such as policy optimisation, including under shifts in environment dynamics. These results have implications for the practical design and selection of data sources for reward learning.

****

PI-GNN: Towards Robust Semi-Supervised Node Classification against Noisy Labels
Xuefeng Du,Tian Bian,Yu Rong,Bo Han,Tongliang Liu,Tingyang Xu,Wenbing Huang,Junzhou Huang

Semi-supervised node classification on graphs is a fundamental problem in graph mining that uses a small set of labeled nodes and many unlabeled nodes for training, so that its performance is quite sensitive to the quality of the node labels. However, it is expensive to maintain the label quality for real-world graph datasets, which presents huge challenges for the learning algorithm to keep a good generalization ability. In this paper, we propose a novel robust learning objective dubbed pairwise interactions (PI) for the model, such as Graph Neural Network (GNN) to combat against noisy labels. Unlike classic robust training approaches that operate on the pointwise interactions between node and class label pairs, PI explicitly forces the embeddings for node pairs that hold a positive PI label to be close to each other, which can be applied to both labeled and unlabeled nodes. We design several instantiations for the PI labels based on the graph structure as well as node class labels, and further propose a new uncertainty-aware training technique to mitigate the negative effect of the sub-optimal PI labels. Extensive experiments on different datasets and GNN architectures demonstrate the effectiveness of PI, which also brings a promising improvement over the state-of-the-art methods.

****

Latent Image Animator: Learning to Animate Images via Latent Space Navigation
Yaohui Wang,Di Yang,Francois Bremond,Antitza Dantcheva

Due to the remarkable progress of deep generative models, animating images has become increasingly efficient, whereas associated results have become increasingly realistic. Current animation-approaches commonly exploit structure representation extracted from driving videos. Such structure representation is instrumental in transferring motion from driving videos to still images. However, such approaches fail in case the source image and driving video encompass large appearance variation. Moreover, the extraction of structure information requires additional modules that endow the animation-model with increased complexity. Deviating from such models, we here introduce the Latent Image Animator (LIA), a self-supervised autoencoder that evades need for structure representation. LIA is streamlined to animate images by linear navigation in the latent space. Specifically, motion in generated video is constructed by linear displacement of codes in the latent space. Towards this, we learn a set of orthogonal motion directions simultaneously, and use their linear combination, in order to represent any displacement in the latent space. Extensive quantitative and qualitative analysis suggests that our model systematically and significantly outperforms state-of-art methods on VoxCeleb, Taichi and TED-talk datasets w.r.t. generated quality.

****

Fully differentiable model discovery
Gert-Jan Both,Remy Kusters

Model discovery aims at autonomously discovering differential equations underlying a dataset. Approaches based on Physics Informed Neural Networks (PINNs) have shown great promise, but a fully-differentiable model which explicitly learns the equation has remained elusive. In this paper we propose such an approach by in

tegrating neural network-based surrogates with Sparse Bayesian Learning (SBL). This combination yields a robust model discovery algorithm, which we showcase on various datasets. We then identify a connection with multitask learning, and build on it to construct a Physics Informed Normalizing Flows (PINFs). We present a proof-of-concept using a PINF to directly learn a density model from single particle data. Our work expands PINNs to various types of neural network architectures, and connects neural network-based surrogates to the rich field of Bayesian parameter inference.
****
D-CODE: Discovering Closed-form ODEs from Observed Trajectories
Zhaozhi Qian,Krzysztof Kacprzyk,Mihaela van der Schaar
For centuries, scientists have manually designed closed-form ordinary differential equations (ODEs) to model dynamical systems. An automated tool to distill closed-form ODEs from observed trajectories would accelerate the modeling process. Traditionally, symbolic regression is used to uncover a closed-form prediction function $a=f(b)$ with label-feature pairs (a_i, b_i) as training examples. However, an ODE models the time derivative $\dot{x}(t)$ of a dynamical system, e.g. $\dot{x}(t) = f(x(t),t)$, and the "label" $\dot{x}(t)$ is usually *not* observed. The existing ways to bridge this gap only perform well for a narrow range of settings with low measurement noise, frequent sampling, and non-chaotic dynamics. In this work, we propose the Discovery of Closed-form ODE framework (D-CODE), which advances symbolic regression beyond the paradigm of supervised learning. D-CODE leverages a novel objective function based on the variational formulation of ODEs to bypass the unobserved time derivative. For formal justification, we prove that this objective is a valid proxy for the estimation error of the true (but unknown) ODE. In the experiments, D-CODE successfully discovered the governing equations of a diverse range of dynamical systems under challenging measurement settings with high noise and infrequent sampling.
****
AdaFocal: Calibration-aware Adaptive Focal Loss
Arindam Ghosh,Thomas Schaaf,Matthew R. Gormley
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 deep networks than cross-entropy loss, while achieving the same level of accuracy \cite{mukhoti2020}. This success stems from Focal loss regularizing the entropy of the network's prediction (controlled by the hyper-parameter γ), thereby reining in the network's overconfidence. Further improvements in calibration can be achieved if γ is selected independently for each training sample. However, the proposed strategy (named FLSD-53) is based on simple heuristics which, when selecting the γ, does not take into account any knowledge of whether the network is under or over confident about such samples and by how much. As a result, in most cases, this strategy performs only slightly better. In this paper, we propose a calibration-aware sample-dependent Focal loss called AdaFocal that adaptively modifies γ from one training step to the next based on the information about the network's current calibration behaviour. At each training step t, AdaFocal adjusts the γ_t based on (1) γ_{t-1} of the previous training step (2) the magnitude of the network's under/over-confidence. We evaluate our proposed method on various image recognition and NLP tasks, covering a variety of network architectures, and confirm that AdaFocal consistently achieves significantly better calibration than the competing state-of-the-art methods without loss of accuracy.
****
Know Thyself: Transferable Visual Control Policies Through Robot-Awareness
Edward S. Hu,Kun Huang,Oleh Rybkin,Dinesh Jayaraman
Training visual control policies from scratch on a new robot typically requires generating large amounts of robot-specific data. How might we leverage data previously collected on another robot to reduce or even completely remove this need for robot-specific data? We propose a "robot-aware control" paradigm that achieves this by exploiting readily available knowledge about the robot. We then insta

ntiate this in a robot-aware model-based RL policy by training modular dynamics models that couple a transferable, robot-aware world dynamics module with a robot-specific, potentially analytical, robot dynamics module. This also enables us to set up visual planning costs that separately consider the robot agent and the world. Our experiments on tabletop manipulation tasks with simulated and real robots demonstrate that these plug-in improvements dramatically boost the transferability of visual model-based RL policies, even permitting zero-shot transfer of visual manipulation skills onto new robots. Project website: https://www.seas.upenn.edu/~hued/rac

****

Node Feature Extraction by Self-Supervised Multi-scale Neighborhood Prediction

Eli Chien,Wei-Cheng Chang,Cho-Jui Hsieh,Hsiang-Fu Yu,Jiong Zhang,Olgica Milenkovic,Inderjit S Dhillon

Learning on graphs has attracted significant attention in the learning community due to numerous real-world applications. In particular, graph neural networks (GNNs), which take \emph{numerical} node features and graph structure as inputs, have been shown to achieve state-of-the-art performance on various graph-related learning tasks. Recent works exploring the correlation between numerical node features and graph structure via self-supervised learning have paved the way for further performance improvements of GNNs. However, methods used for extracting numerical node features from \emph{raw data} are still \emph{graph-agnostic} within standard GNN pipelines. This practice is sub-optimal as it prevents one from fully utilizing potential correlations between graph topology and node attributes. To mitigate this issue, we propose a new self-supervised learning framework, Graph Information Aided Node feature exTraction (GIANT). GIANT makes use of the eXtreme Multi-label Classification (XMC) formalism, which is crucial for fine-tuning the language model based on graph information, and scales to large datasets. We also provide a theoretical analysis that justifies the use of XMC over link prediction and motivates integrating XR-Transformers, a powerful method for solving XMC problems, into the GIANT framework. We demonstrate the superior performance of GIANT over the standard GNN pipeline on Open Graph Benchmark datasets: For example, we improve the accuracy of the top-ranked method GAMLP from 68.25% to 69.67%, SGC from 63.29% to 66.10% and MLP from 47.24% to 61.10% on the ogbn-papers100M dataset by leveraging GIANT.

****

Prioritized training on points that are learnable, worth learning, and not yet learned

Sören Mindermann,Muhammed Razzak,Mrinank Sharma,Jan M. Brauner,Winnie Xu,Andreas Kirsch,Aidan Gomez,Benedikt Höltgen,Sebastian Farquhar,Yarin Gal

We introduce reducible held-out loss selection (RHOLS), a technique for faster model training which selects a sequence of training points that are “just right”. We propose a tractable information-theoretic acquisition function—the reducible heldout loss—to efficiently choose training points that maximize information about a holdout set. We show that the “hard” (e.g. high loss) points usually selected in the optimization literature are typically noisy, leading to deterioration on real-world datasets. At the same time, “easy” (e.g. low noise) samples, often prioritized for curriculum learning, confer less information. In contrast, RHOLS chooses points that are “just right” and trains in fewer steps than the above approaches.

****

Spherical Message Passing for 3D Molecular Graphs

Yi Liu,Limei Wang,Meng Liu,Yuchao Lin,Xuan Zhang,Bora Oztekin,Shuiwang Ji

We consider representation learning of 3D molecular graphs in which each atom is associated with a spatial position in 3D. This is an under-explored area of research, and a principled message passing framework is currently lacking. In this work, we conduct analyses in the spherical coordinate system (SCS) for the complete identification of 3D graph structures. Based on such observations, we propose the spherical message passing (SMP) as a novel and powerful scheme for 3D molecular learning. SMP dramatically reduces training complexity, enabling it to perform efficiently on large-scale molecules. In addition, SMP is capable of distin guishing almost all molecular structures, and the uncovered cases may not exist in practice. Based on meaningful physically-based representations of 3D information, we further propose the SphereNet for 3D molecular learning. Experimental results demonstrate that the use of meaningful 3D information in SphereNet leads to significant performance improvements in prediction tasks. Our results also demonstrate the advantages of SphereNet in terms of capability, efficiency, and scalability.

****

Fairness Guarantees under Demographic Shift

Stephen Giguere,Blossom Metevier,Bruno Castro da Silva,Yuriy Brun,Philip S. Thomas,Scott Niekum

Recent studies have demonstrated that using machine learning for social applications can lead to injustice in the form of racist, sexist, and otherwise unfair and discriminatory outcomes. To address this challenge, recent machine learning algorithms have been designed to limit the likelihood such unfair behaviors will occur. However, these approaches typically assume the data used for training is representative of what will be encountered once the model is deployed, thus limiting their usefulness. In particular, if certain subgroups of the population become more or less probable after the model is deployed (a phenomenon we call demographic shift), the fair-ness assurances provided by prior algorithms are often invalid. We consider the impact of demographic shift and present a class of algorithms, called Shifty algorithms, that provide high-confidence behavioral guarantees that hold under demographic shift. Shifty is the first technique of its kind and demonstrates an effective strategy for designing algorithms to overcome the challenges demographic shift poses. We evaluate Shifty-ttest, an implementation of Shifty based on Student’s ■-test, and, using a real-world data set of university entrance exams and subsequent student success, show that the models output by our algorithm avoid unfair bias under demo-graphic shift, unlike existing methods. Our experiments demonstrate that our algorithm’s high-confidence fairness guarantees are valid in practice and that our algorithm is an effective tool for training models that are fair when demographic shift occurs.

****

Spanning Tree-based Graph Generation for Molecules

Sungsoo Ahn,Binghong Chen,Tianzhe Wang,Le Song

In this paper, we explore the problem of generating molecules using deep neural networks, which has recently gained much interest in chemistry. To this end, we propose a spanning tree-based graph generation (STGG) framework based on formulating molecular graph generation as a construction of a spanning tree and the residual edges. Such a formulation exploits the sparsity of molecular graphs and allows using compact tree-constructive operations to define the molecular graph connectivity. Based on the intermediate graph structure of the construction process, our framework can constrain its generation to molecular graphs that satisfy the chemical valence rules. We also newly design a Transformer architecture with tree-based relative positional encodings for realizing the tree construction procedure. Experiments on QM9, ZINC250k, and MOSES benchmarks verify the effectiveness of the proposed framework in metrics such as validity, Frechet ChemNet distance, and fragment similarity. We also demonstrate the usefulness of STGG in maximizing penalized LogP value of molecules.

****

Predicting subscriber usage: Analyzing multi-dimensional time-series using Convolutional Neural Networks

Benjamin Azaria,Lee-Ad Gottlieb

Companies operating under the subscription model typically invest significant resources attempting to predict customer's feature usage. These predictions can be used to fuel growth: It may allow these companies to target individual customers -- for example to convert non-paying consumers to begin paying for for enhanced services -- or to identify customers not maximizing their subscription product.

This assistance can avoid an increase in the churn rate, and for some consumers may increase their usage.

In this work, we develop a deep learning model to predict the product usage of a given consumer, based on historical usage. We adapt a Convolutional Neural Network to time-series data followed by Auxiliary Output, and demonstrate that this enhanced model effectively predicts future change in usage.

****

Policy improvement by planning with Gumbel

Ivo Danihelka,Arthur Guez,Julian Schrittwieser,David Silver

AlphaZero is a powerful reinforcement learning algorithm based on approximate policy iteration and tree search. However, AlphaZero can fail to improve its policy network, if not visiting all actions at the root of a search tree. To address this issue, we propose a policy improvement algorithm based on sampling actions without replacement. Furthermore, we use the idea of policy improvement to replace the more heuristic mechanisms by which AlphaZero selects and uses actions, both at root nodes and at non-root nodes. Our new algorithms, Gumbel AlphaZero and Gumbel MuZero, respectively without and with model-learning, match the state of the art on Go, chess, and Atari, and significantly improve prior performance when planning with few simulations.

****

Fooling Explanations in Text Classifiers

Adam Ivankay,Ivan Girardi,Chiara Marchiori,Pascal Frossard

State-of-the-art text classification models are becoming increasingly reliant on deep neural networks (DNNs). Due to their black-box nature, faithful and robust explanation methods need to accompany classifiers for deployment in real-life scenarios. However, it has been shown that explanation methods in vision applications are susceptible to local, imperceptible perturbations that can significantly alter the explanations without changing the predicted classes. We show here that the existence of such perturbations extends to text classifiers as well. Specifically, we introduce TextExplanationFooler (TEF), a novel explanation attack algorithm that alters text input samples imperceptibly so that the outcome of widely-used explanation methods changes considerably while leaving classifier predictions unchanged. We evaluate the attribution robustness estimation performance of TEF on five text classification datasets, utilizing three DNN architectures and a transformer architecture for each dataset. By significantly decreasing the correlation between unchanged and perturbed input attributions, we show that all models and explanation methods are susceptible to TEF perturbations. Moreover, we evaluate how the perturbations transfer to other model architectures and attribution methods, finding better than random performance in scenarios where the exact attacked model and explanation method are unknown. Finally, we introduce a semi-universal attack that is able to compute fast, computationally light perturbations with no knowledge of the attacked classifier nor explanation method. Overall, our work shows that explanations in text classifiers are fragile and users need to carefully address their robustness before relying on them in critical applications.

****

On the Learning and Learnability of Quasimetrics

Tongzhou Wang,Phillip Isola

Our world is full of asymmetries. Gravity and wind can make reaching a place easier than coming back. Social artifacts such as genealogy charts and citation graphs are inherently directed. In reinforcement learning and control, optimal goal-reaching strategies are rarely reversible (symmetrical). Distance functions supported on these asymmetrical structures are called quasimetrics. Despite their common appearance, little research has been done on the learning of quasimetrics. Our theoretical analysis reveals that a common class of learning algorithms, including unconstrained multilayer perceptrons (MLPs), provably fails to learn a quasimetric consistent with training data. In contrast, our proposed Poisson Quasimetric Embedding (PQE) is the first quasimetric learning formulation that both is learnable with gradient-based optimization and enjoys strong performance guarantees. Experiments on random graphs, social graphs, and offline Q-learning demonstrate its effectiveness over many common baselines.

****

End-to-End Balancing for Causal Continuous Treatment-Effect Estimation

Mohammad Taha Bahadori,Eric Tchetgen Tchetgen,David Heckerman

We study the problem of observational causal inference with continuous treatment. We focus on the challenge of estimating the causal response curve for infrequently-observed treatment values.

We design a new algorithm based on the framework of entropy balancing which learns weights that directly maximize causal inference accuracy using end-to-end optimization. Our weights can be customized for different datasets and causal inference algorithms. We propose a new theory for consistency of entropy balancing for continuous treatments. Using synthetic and real-world data, we show that our proposed algorithm outperforms the entropy balancing in terms of causal inference accuracy.

****

Inference-Time Personalized Federated Learning

Ohad Amosy,Gal Eyal,Gal Chechik

In Federated learning (FL), multiple clients collaborate to learn a model through a central server but keep the data decentralized. Personalized federated learning (PFL) further extends FL to handle data heterogeneity between clients by learning personalized models. In both FL and PFL, all clients participate in the training process and their labeled data is used for training. However, in reality, novel clients may wish to join a prediction service after it has been deployed, obtaining predictions for their own unlabeled data.

Here, we defined a new learning setup, Inference-Time PFL (IT-PFL), where a model trained on a set of clients, needs to be later evaluated on novel unlabeled clients at inference time. We propose a novel approach to this problem IT-PFL-HN, based on a hypernetwork module and an encoder module. Specifically, we train an encoder network that learns a representation for a client given its unlabeled data. That client representation is fed to a hypernetwork that generates a personalized model for that client. Evaluated on four benchmark datasets, we find that IT-PFL-HN generalizes better than current FL and PFL methods, especially when the novel client has a large domain shift. We also analyzed the generalization error for the novel client, showing how it can be bounded using results from multi-task learning and domain adaptation. Finally, since novel clients do not contribute their data to training, they can potentially have better control over their data privacy; Indeed, we showed analytically and experimentally how novel clients can apply differential privacy to their data.

****

Data-Efficient Augmentation for Training Neural Networks

Tian Yu Liu,Baharan Mirzasoleiman

Data augmentation is essential to achieve state-of-the-art performance in many deep learning applications. However, modern data augmentation techniques become computationally prohibitive for large 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, speeds up learning by enlarging the smaller singular values of the network Jacobian. 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 label/residual vector. 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 max-loss strategy by 7.7% on CIFAR10 while achieving 6.3x speedup, and by 4.7% on SVHN while achieving 2.2x speedup, using 10% and 30% subsets, respectively.

****

On The Quality Assurance Of Concept-Based Representations

Mateo Espinosa Zarlenga,Pietro Barbiero,Zohreh Shams,Dmitry Kazhdan,Umang Bhatt,Mateja Jamnik

Recent work on Explainable AI has focused on concept-based explanations, where d eep learning models are explained in terms of high-level units of information, referred to as concepts. In parallel, the field of disentanglement learning has explored the related notion of finding underlying factors of variation in the data that have interpretability properties. Despite their overlapping purpose, the metrics to evaluate the quality of concepts and factors of variation in the two fields are not aligned, hindering a systematic comparison. In this paper we consider factors of variation as concepts and thus unify the notations in concept and disentanglement learning. Next, we propose metrics for evaluating the quality of concept representations in both approaches, in the presence and in the absence of ground truth concept labels. Via our proposed metrics, we benchmark state-of-the-art methods from both families, and propose a set of guidelines to determine the impact that supervision may have on the quality of learnt concept representations.

****

Learning Optimal Conformal Classifiers

David Stutz,Krishnamurthy Dj Dvijotham,Ali Taylan Cemgil,Arnaud Doucet

Modern deep learning based classifiers show very high accuracy on test data but this does not provide sufficient guarantees for safe deployment, especially in high-stake AI applications such as medical diagnosis. Usually, predictions are obtained without a reliable uncertainty estimate or a formal guarantee. Conformal prediction (CP) addresses these issues by using the classifier's predictions, e.g., its probability estimates, to predict confidence sets containing the true class with a user-specified probability. However, using CP as a separate processing step after training prevents the underlying model from adapting to the prediction of confidence sets. Thus, this paper explores strategies to differentiate through CP during training with the goal of training model with the conformal wrapper end-to-end. In our approach, conformal training (ConfTr), we specifically "simulate" conformalization on mini-batches during training. Compared to standard training, ConfTr reduces the average confidence set size (inefficiency) of state-of-the-art CP methods applied after training. Moreover, it allows to "shape" the confidence sets predicted at test time, which is difficult for standard CP. On experiments with several datasets, we show ConfTr can influence how inefficiency is distributed across classes, or guide the composition of confidence sets in terms of the included classes, while retaining the guarantees offered by CP.

****

Learning Prototype-oriented Set Representations for Meta-Learning

Dan dan Guo,Long Tian,Minghe Zhang,Mingyuan Zhou,Hongyuan Zha

Learning from set-structured data is a fundamental problem that has recently attracted increasing attention, where a series of summary networks are introduced to deal with the set input. In fact, many meta-learning problems can be treated as set-input tasks. Most existing summary networks aim to design different architectures for the input set in order to enforce permutation invariance. However, scant attention has been paid to the common cases where different sets in a meta distribution are closely related and share certain statistical properties. Viewing each set as a distribution over a set of global prototypes, this paper provides a novel prototype-oriented optimal transport (POT) framework to improve existing summary networks. To learn the distribution over the global prototypes, we minimize its regularized optimal transport distance to the set empirical distribution over data points, providing a natural unsupervised way to improve the summary network. Since our plug-and-play framework can be applied to many meta learning problems, we further instantiate it to the cases of few-shot classification and implicit meta generative modeling. Extensive experiments demonstrate that our framework significantly improves the existing summary networks on learning more powerful summary statistics from sets and can be successfully integrated into metric-based few-shot classification and generative modeling applications, providing a promising tool for addressing set-input and meta-learning problems.

****

Universality of Deep Neural Network Lottery Tickets: A Renormalization Group Perspective

William T Redman,Tianlong Chen,Akshunna S. Dogra,Zhangyang Wang

Foundational work on the Lottery Ticket Hypothesis has suggested an exciting corollary: winning tickets found in the context of one task can be transferred to similar tasks, possibly even across different architectures. While this has become of broad practical and theoretical interest, to date, there exists no detailed understanding of why winning ticket universality exists, or any way of knowing a priori whether a given ticket can be transferred to a given task. To address these outstanding open questions, we make use of renormalization group theory, one of the most successful tools in theoretical physics. We find that iterative magnitude pruning, the method used for discovering winning tickets, is a renormalization group scheme. This opens the door to a wealth of existing numerical and theoretical tools, some of which we leverage here to examine winning ticket universality in large scale lottery ticket experiments, as well as sheds new light on the success iterative magnitude pruning has found in the field of sparse machine learning.

****

Interpreting Reinforcement Policies through Local Behaviors

Ronny Luss,Amit Dhurandhar,Miao Liu

Many works in explainable AI have focused on explaining black-box classification models. Explaining deep reinforcement learning (RL) policies in a manner that could be understood by domain users has received much less attention. In this paper, we propose a novel perspective to understanding RL policies based on identifying important states from automatically learned meta-states. The key conceptual difference between our approach and many previous ones is that we form meta-states based on locality governed by the expert policy dynamics rather than based on similarity of actions, and that we do not assume any particular knowledge of the underlying topology of the state space. Theoretically, we show that our algorithm to find meta-states converges and the objective that selects important states from each meta-state is submodular leading to efficient high quality greedy selection. Experiments on three domains (four rooms, door-key and minipacman) and a carefully conducted user study illustrate that our perspective leads to better understanding of the policy. We conjecture that this is a result of our meta-states being more intuitive in that the corresponding important states are strong indicators of tractable intermediate goals that are easier for humans to interpret and follow.

****

Accuracy-Privacy Trade-off in Deep Ensemble: A Membership Inference Perspective

Shahbaz Rezaei,Zubair Shafiq,Xin Liu

Deep ensemble learning has been shown to improve accuracy by training multiple neural networks and fusing their outputs. Ensemble learning has also been used to defend against membership inference attacks that undermine privacy. In this paper, we empirically demonstrate a trade-off between these two goals, namely accuracy and privacy (in terms of membership inference attacks), in deep ensembles. Using a wide range of datasets and model architectures, we show that the effectiveness of membership inference attacks also increases when ensembling improves accuracy. To better understand this trade-off, we study the impact of various factors such as prediction confidence and agreement between models that constitute the ensemble. Finally, we evaluate defenses against membership inference attacks based on regularization and differential privacy. We show that while these defenses can mitigate the effectiveness of the membership inference attack, they simultaneously degrade ensemble accuracy. We illustrate similar trade-off in more advanced and state-of-the-art ensembling techniques, such as snapshot ensembles and diversified ensemble networks. The source code is available in supplementary materials.

****

Neuro-Symbolic Forward Reasoning

Hikaru Shindo,Devendra Singh Dhami,Kristian Kersting

Reasoning is an essential part of human intelligence and thus has been a long-standing goal in artificial intelligence research. With the recent success of deep learning, incorporating reasoning with deep learning systems i.e. neuro-symbolic AI has become a major field of interest. We propose Neuro-Symbolic Forward Rea

soner (NS-FR), a new approach for reasoning tasks taking advantage of differentiable forward-chaining using first-order logic. The key idea is to combine differentiable forward-chaining reasoning with object-centric learning. Differentiable forward-chaining reasoning computes logical entailments smoothly, i.e., it deduces new facts from given facts and rules in a differentiable manner. The object-centric learning approach factorizes raw inputs into representations in terms of objects. This allows us to provide a consistent framework to perform the forward-chaining inference from raw inputs. NS-FR factorizes the raw inputs into the object-centric representations, then converts them into probabilistic ground atoms and finally performs differentiable forward-chaining inference using weighted rules for inference. Our comprehensive experimental evaluations on object-centric reasoning data sets, 2D Kandinsky patterns and 3D CLEVR-Hans, and variety of tasks show the effectiveness and advantage of our approach.
****

Embedded-model flows: Combining the inductive biases of model-free deep learning and explicit probabilistic modeling
Gianluigi Silvestri,Emily Fertig,Dave Moore,Luca Ambrogioni
Normalizing flows have shown great success as general-purpose density estimators. However, many real world applications require the use of domain-specific knowledge, which normalizing flows cannot readily incorporate. We propose embedded-model flows (EMF), which alternate general-purpose transformations with structured layers that embed domain-specific inductive biases. These layers are automatically constructed by converting user-specified differentiable probabilistic models into equivalent bijective transformations. We also introduce gated structured layers, which allow bypassing the parts of the models that fail to capture the statistics of the data. We demonstrate that EMFs can be used to induce desirable properties such as multimodality and continuity. Furthermore, we show that EMFs enable a high performance form of variational inference where the structure of the prior model is embedded in the variational architecture. In our experiments, we show that this approach outperforms a large number of alternative methods in common structured inference problems.
****

On the Expressiveness and Learning of Relational Neural Networks on Hypergraphs
Zhezheng Luo,Jiayuan Mao,Joshua B. Tenenbaum,Leslie Pack Kaelbling
This paper presents a framework for analyzing the expressiveness and learning of relational models applied to hypergraph reasoning tasks. We start with a general framework that unifies several relational neural network architectures: graph neural networks, neural logical machines, and transformers. Our first contribution is a fine-grained analysis of the expressiveness of these neural networks, that is, the set of functions that they can realize and the set of problems that they can solve. Our result is a hierarchy of problems they can solve, defined in terms of various hyperparameters such as depth and width. Next, we analyze the learning properties of these neural networks, especially focusing on how they can be trained on a small graphs and generalize to larger graphs. Our theoretical results are further supported by the empirical results illustrating the optimization and generalization of these models based on gradient-descent training.
****

Neural Architecture Search via Ensemble-based Knowledge Distillation
Fanxin Li,Shixiong Zhao,Haowen Pi,Yuhao QING,Yichao Fu,Sen Wang,Heming Cui
Neural Architecture Search (NAS) automatically searches for well-performed network architectures from a given search space. The One-shot NAS method improves the training efficiency by sharing weights among the possible architectures in the search space, but unfortunately suffers from insufficient parameterization of each architecture due to interferences from other architectures. Recent works attempt to alleviate the insufficient parameterization problem by knowledge distillation, which let the learning of all architectures (students) be guided by the knowledge (i.e., parameters) from a better-parameterized network (teacher), which can be either a pre-trained one (e.g., ResNet50) or some searched out networks with good accuracy performance up to now.

However, all these methods fall short in providing a sufficiently outstanding teacher, as they either depend on a pre-trained network that does not fit the NAS task the best, or the selected fitting teachers are still undertrained and inaccurate. In this paper, we take the first step to propose an ensemble-based knowledge distillation method for NAS, called EnNAS, which assembles an outstanding teacher by aggregating a set of architectures currently searched out with the most diversity (high diversity brings highly accurate ensembles); by doing so, EnNAS can deliver a high-quality knowledge distillation with outstanding teacher network (i.e., the ensemble network) all the time. Eventually, compared with existing works, on the real-world dataset ImageNet, EnNAS improved the top-1 accuracy of architectures searched out by 1.2% on average and 3.3% at most.

****

Multitask Prompted Training Enables Zero-Shot Task Generalization

Victor Sanh,Albert Webson,Colin Raffel,Stephen Bach,Lintang Sutawika,Zaid Alyafeai,Antoine Chaffin,Arnaud Stiegler,Arun Raja,Manan Dey,M Saiful Bari,Canwen Xu,Urmish Thakker,Shanya Sharma Sharma,Eliza Szczechla,Taewoon Kim,Gunjan Chhablani,Nihal Nayak,Debajyoti Datta,Jonathan Chang,Mike Tian-Jian Jiang,Han Wang,Matteo Manica,Sheng Shen,Zheng Xin Yong,Harshit Pandey,Rachel Bawden,Thomas Wang,Trishala Neeraj,Jos Rozen,Abheesht Sharma,Andrea Santilli,Thibault Fevry,Jason Alan Fries,Ryan Teehan,Teven Le Scao,Stella Biderman,Leo Gao,Thomas Wolf,Alexander M Rush

Large language models have recently been shown to attain reasonable zero-shot generalization on a diverse set of tasks (Brown et al., 2020). It has been hypothesized that this is a consequence of implicit multitask learning in language models’ pretraining (Radford et al., 2019). Can zero-shot generalization instead be directly induced by explicit multitask learning? To test this question at scale, we develop a system for easily mapping any natural language tasks into a human-readable prompted form. We convert a large set of supervised datasets, each with multiple prompts with diverse wording. These prompted datasets allow for benchmarking the ability of a model to perform completely unseen tasks specified in natural language. We fine-tune a pretrained encoder-decoder model (Raffel et al., 2020; Lester et al., 2021) on this multitask mixture covering a wide variety of tasks. The model attains strong zero-shot performance on several datasets, often outperforming models 16× its size. Further, our model attains strong performance on a subset of tasks from the BIG-Bench benchmark, outperforming models 6× its size. All trained models are available at https://github.com/bigscience-workshop/t-zero, and all prompts are available at https://github.com/bigscience-workshop/promptsource.

****

Learning Audio-Visual Dereverberation

Changan Chen,Wei Sun,David Harwath,Kristen Grauman

Reverberation from audio reflecting off surfaces and objects in the environment not only degrades the quality of speech for human perception, but also severely impacts the accuracy of automatic speech recognition. Prior work attempts to remove reverberation based on the audio modality only. Our idea is to learn to dereverberate speech from audio-visual observations. The visual environment surrounding a human speaker reveals important cues about the room geometry, materials, and speaker location, all of which influence the precise reverberation effects in the audio stream. We introduce Visually-Informed Dereverberation of Audio (VIDA), an end-to-end approach that learns to remove reverberation based on both the observed sounds and visual scene. In support of this new task, we develop a large-scale dataset that uses realistic acoustic renderings of speech in real-world 3D scans of homes offering a variety of room acoustics. Demonstrating our approach on both simulated and real imagery for speech enhancement, speech recognition, and speaker identification, we show it achieves state-of-the-art performance and substantially improves over traditional audio-only methods.

****

A Relational Intervention Approach for Unsupervised Dynamics Generalization in Model-Based Reinforcement Learning

Jiaxian Guo,Mingming Gong,Dacheng Tao

The generalization of model-based reinforcement learning (MBRL) methods to environments with unseen transition dynamics is an important yet challenging problem. Existing methods try to extract environment-specified information Z from past transition segments to make the dynamics prediction model generalizable to different dynamics. However, because environments are not labelled, the extracted information inevitably contains redundant information unrelated to the dynamics in transition segments and thus fails to maintain a crucial property of Z: Z should be similar in the same environment and dissimilar in different ones. As a result, the learned dynamics prediction function will deviate from the true one, which undermines the generalization ability. To tackle this problem, we introduce an interventional prediction module to estimate the probability of two estimated $\hat{z}_i, \hat{z}_j$ belonging to the same environment.
Furthermore, by utilizing the Z's invariance within a single environment, a relational head is proposed to enforce the similarity between $\hat{{Z}}$ from the same environment. As a result, the redundant information will be reduced in $\hat{Z}$. We empirically show that $\hat{{Z}}$ estimated by our method enjoy less redundant information than previous methods, and such $\hat{{Z}}$ can significantly reduce dynamics prediction errors and improve the performance of model-based RL methods on zero-shot new environments with unseen dynamics. The codes of this method are available at \url{https://github.com/CR-Gjx/RIA}.
****
Aug-ILA: More Transferable Intermediate Level Attacks with Augmented References
Chiu Wai Yan,Dit-Yan Yeung
An intriguing property of deep neural networks is that adversarial attacks can transfer across different models. Existing methods such as the Intermediate Level Attack (ILA) further improve black-box transferability by fine-tuning a reference adversarial attack, so as to maximize the perturbation on a pre-specified layer of the source model. In this paper, we revisit ILA and evaluate the effect of applying augmentation to the images before passing them to ILA. We start by looking into the effect of common image augmentation techniques and exploring novel augmentation with the aid of adversarial perturbations. Based on the observations, we propose Aug-ILA, an improved method that enhances the transferability of an existing attack under the ILA framework. Specifically, Aug-ILA has three main characteristics: typical image augmentation such as random cropping and resizing applied to all ILA inputs, reverse adversarial update on the clean image, and interpolation between two attacks on the reference image. Our experimental results show that Aug-ILA outperforms ILA and its subsequent variants, as well as state-of-the-art transfer-based attacks, by achieving 96.99% and 87.84% average attack success rates with perturbation budgets 0.05 and 0.03, respectively, on nine undefended models.
****
Bayesian Relational Generative Model for Scalable Multi-modal Learning
Ehsan Hajiramezanali,Talip Ucar,Lindsay Edwards
The study of complex systems requires the integration of multiple heterogeneous and high-dimensional data types (e.g. multi-omics). However, previous generative approaches for multi-modal inputs suffer from two shortcomings. First, they are not stochastic processes, leading to poor uncertainty estimations over their predictions. This is mostly due to the computationally intensive nature of traditional stochastic processes, such as Gaussian Processes (GPs), that makes their applicability limited in multi-modal learning frameworks. Second, they are not able to effectively approximate the joint posterior distribution of multi-modal data types with various missing patterns. More precisely, their model assumptions result in miscalibrated precisions and/or computational cost of sub-sampling procedure. In this paper, we propose a class of stochastic processes that learns a graph of dependencies between samples across multi-modal data types through adopting priors over the relational structure of the given data modalities. The dependency graph in our method, multi-modal Relational Neural Process (mRNP), not only posits distributions over the functions and naturally enables rapid adaptation to new observations by its predictive distribution, but also makes mRNP scalable to large datasets through mini-batch optimization. We also introduce mixture-

of-graphs (MoG) in our model construction and show that it can address the aforementioned limitations in joint posterior approximation. Experiments on both toy regression and classification tasks using real-world datasets demonstrate the potential of mRNP for offering higher prediction accuracies as well as more robust uncertainty estimates compared to existing baselines and state-of-the-art methods.

****

Continuous-Time Meta-Learning with Forward Mode Differentiation
Tristan Deleu,David Kanaa,Leo Feng,Giancarlo Kerg,Yoshua Bengio,Guillaume Lajoie,Pierre-Luc Bacon

Drawing inspiration from gradient-based meta-learning methods with infinitely small gradient steps, we introduce Continuous-Time Meta-Learning (COMLN), a meta-learning algorithm where adaptation follows the dynamics of a gradient vector field. Specifically, representations of the inputs are meta-learned such that a task-specific linear classifier is obtained as a solution of an ordinary differential equation (ODE). Treating the learning process as an ODE offers the notable advantage that the length of the trajectory is now continuous, as opposed to a fixed and discrete number of gradient steps. As a consequence, we can optimize the amount of adaptation necessary to solve a new task using stochastic gradient descent, in addition to learning the initial conditions as is standard practice in gradient-based meta-learning. Importantly, in order to compute the exact meta-gradients required for the outer-loop updates, we devise an efficient algorithm based on forward mode differentiation, whose memory requirements do not scale with the length of the learning trajectory, thus allowing longer adaptation in constant memory. We provide analytical guarantees for the stability of COMLN, we show empirically its efficiency in terms of runtime and memory usage, and we illustrate its effectiveness on a range of few-shot image classification problems.

****

Non-reversible Parallel Tempering for Uncertainty Approximation in Deep Learning
Wei Deng,Qian Zhang,Qi Feng,Faming Liang,Guang Lin

Parallel tempering (PT), also known as replica exchange, is the go-to workhorse for simulations of multi-modal distributions. The key to the success of PT is to adopt efficient swap schemes. The popular deterministic even-odd (DEO) scheme exploits the non-reversibility property and has successfully reduced the communication cost from $O(P^2)$ to $O(P)$ given sufficient many P chains. However, such an innovation largely disappears given limited chains in big data problems due to the extremely few bias-corrected swaps. To handle this issue, we generalize the DEO scheme to promote the non-reversibility and obtain an optimal communication cost $O(P\log P)$. In addition, we also analyze the bias when we adopt stochastic gradient descent (SGD) with large and constant learning rates as exploration kernels. Such a user-friendly nature enables us to conduct large-scale uncertainty approximation tasks without much tuning costs.

****

Critical Points in Quantum Generative Models
Eric Ricardo Anschuetz

One of the most important properties of neural networks is the clustering of local minima of the loss function near the global minimum, enabling efficient training. Though generative models implemented on quantum computers are known to be more expressive than their traditional counterparts, it has empirically been observed that these models experience a transition in the quality of their local minima. Namely, below some critical number of parameters, all local minima are far from the global minimum in function value; above this critical parameter count, all local minima are good approximators of the global minimum. Furthermore, for a certain class of quantum generative models, this transition has empirically been observed to occur at parameter counts exponentially large in the problem size, meaning practical training of these models is out of reach. Here, we give the first proof of this transition in trainability, specializing to this latter class of quantum generative model. We use techniques inspired by those used to study the loss landscapes of classical neural networks. We also verify that our analytic results hold experimentally even at modest model sizes.

****

VOS: Learning What You Don't Know by Virtual Outlier Synthesis

Xuefeng Du,Zhaoning Wang,Mu Cai,Yixuan Li

Out-of-distribution (OOD) detection has received much attention lately due to its importance in the safe deployment of neural networks. One of the key challenges is that models lack supervision signals from unknown data, and as a result, can produce overconfident predictions on OOD data. Previous approaches rely on real outlier datasets for model regularization, which can be costly and sometimes infeasible to obtain in practice. In this paper, we present VOS, a novel framework for OOD detection by adaptively synthesizing virtual outliers that can meaningfully regularize the model's decision boundary during training. Specifically, VOS samples virtual outliers from the low-likelihood region of the class-conditional distribution estimated in the feature space. Alongside, we introduce a novel unknown-aware training objective, which contrastively shapes the uncertainty space between the ID data and synthesized outlier data. VOS achieves competitive performance on both object detection and image classification models, reducing the FPR95 by up to 9.36% compared to the previous best method on object detectors. Code is available at https://github.com/deeplearning-wisc/vos.

****

Spending Thinking Time Wisely: Accelerating MCTS with Virtual Expansions

Weirui Ye,Pieter Abbeel,Yang Gao

One of the most important AI research questions is to trade off computation versus performance, since "perfect rational" exists in theory but it is impossible to achieve in practice. Recently, Monte-Carlo tree search (MCTS) has attracted considerable attention due to the significant improvement of performance in varieties of 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 mimics the human behavior that spends adequate amounts of time to think about different questions. Inspired by this, we propose a strategy that converges to the ground truth MCTS search results with much less computation. We give theoretical bounds of the V-MCTS and evaluate the performance in 9×9 Go board games and Atari games. Experiments show that our method can achieve similar performances as the original search algorithm while requiring less than 50% number of search times on average.

We believe that this approach is a viable alternative for tasks with limited time and resources.

****

A framework of deep neural networks via the solution operator of partial differential equations

Wenqi Tao,Zuoqiang Shi

There is a close connection between deep neural networks (DNN) and partial differential equations (PDEs). Many DNN architectures can be modeled by PDEs and have been proposed in the literature. However, their neural network design space is restricted due to the specific form of PDEs, which prevents the design of more effective neural network structures. In this paper, we attempt to derive a general form of PDEs for the design of ResNet-like DNN. To achieve this goal, we first formulate DNN as an adjustment operator applied on the base classifier. Then based on several reasonable assumption, we show the adjustment operator for ResNet-like DNN is the solution operator of PDEs. To show the effectiveness for general form of PDEs, we show that several effective networks can be interpreted by our general form of PDEs and design a training method motivated by PDEs theory to train DNN models for better robustness and less chance of overfitting. Theoretically, we prove that the robustness of DNN trained with our method is certifiable and our training method reduces the generalization gap for DNN. Furthermore, we demonstrate that DNN trained with our method can achieve better generalization and is more resistant to adversarial perturbations than baseline model.

****

Trust Region Policy Optimisation in Multi-Agent Reinforcement Learning

Jakub Grudzien Kuba,Ruiqing Chen,Muning Wen,Ying Wen,Fanglei Sun,Jun Wang,Yaodong Yang

Trust region methods rigorously enabled reinforcement learning (RL) agents to learn monotonically improving policies, leading to superior performance on a variety of tasks. Unfortunately, when it comes to multi-agent reinforcement learning (MARL), the property of monotonic improvement may not simply apply; this is because agents, even in cooperative games, could have conflicting directions of policy updates. As a result, achieving a guaranteed improvement on the joint policy where each agent acts individually remains an open challenge. In this paper, we extend the theory of trust region learning to MARL. Central to our findings are the multi-agent advantage decomposition lemma and the sequential policy update scheme. Based on these, we develop Heterogeneous-Agent Trust Region Policy Optimisation (HATPRO) and Heterogeneous-Agent Proximal Policy Optimisation (HAPPO) algorithms. Unlike many existing MARL algorithms, HATRPO/HAPPO do not need agents to share parameters, nor do they need any restrictive assumptions on decomposibility of the joint value function. Most importantly, we justify in theory the monotonic improvement property of HATRPO/HAPPO. We evaluate the proposed methods on a series of Multi-Agent MuJoCo and StarCraftII tasks. Results show that HATRPO and HAPPO significantly outperform strong baselines such as IPPO, MAPPO and MADDPG on all tested tasks, thereby establishing a new state of the art.
****

Semi-supervised learning of partial differential operators and dynamical flows
Michael Rotman,Amit Dekel,Ran Ilan Ber,Lior Wolf,Yaron Oz

The evolution of dynamical systems is generically governed by nonlinear partial differential equations (PDEs), whose solution, in a simulation framework, requires vast amounts of computational resources. For a growing number of specific cases, neural network-based solvers have been shown to provide comparable results to other numerical methods while utilizing fewer resources.

In this work, we present a novel method that combines a hyper-network solver with a Fourier Neural Operator architecture. Our method treats time and space separately. As a result, it successfully propagates initial conditions in discrete time steps by employing the general composition properties of the partial differential operators. Following previous work, supervision is provided at a specific time point. We test our method on various time evolution PDEs, including nonlinear fluid flows in one, two, and three spatial dimensions. The results show that the new method improves the learning accuracy at the time point of supervision point, and is also able to interpolate and extrapolate the solutions to arbitrary times.
****

Unsupervised Disentanglement with Tensor Product Representations on the Torus
Michael Rotman,Amit Dekel,Shir Gur,Yaron Oz,Lior Wolf

The current methods for learning representations with auto-encoders almost exclusively employ vectors as the latent representations. In this work, we propose to employ a tensor product structure for this purpose. This way, the obtained representations are naturally disentangled. In contrast to the conventional variations methods, which are targeted toward normally distributed features, the latent space in our representation is distributed uniformly over a set of unit circles. We argue that the torus structure of the latent space captures the generative factors effectively. We employ recent tools for measuring unsupervised disentanglement, and in an extensive set of experiments demonstrate the advantage of our method in terms of disentanglement, completeness, and informativeness. The code for our proposed method is available at https://github.com/rotmanmi/Unsupervised-Disentanglement-Torus.
****

Anomaly Detection for Tabular Data with Internal Contrastive Learning
Tom Shenkar,Lior Wolf

We consider the task of finding out-of-class samples in tabular data, where little can be assumed on the structure of the data. In order to capture the structure of the samples of the single training class, we learn mappings that maximize the mutual information between each sample and the part that is masked out. The mappings are learned by employing a contrastive loss, which considers only one sample at a time. Once learned, we can score a test sample by measuring whether t he learned mappings lead to a small contrastive loss using the masked parts of this sample. Our experiments show that our method leads by a sizable accuracy gap in comparison to the literature and that the same default set of hyperparameters provides state-of-the-art results across benchmarks.

****

DESTA: A Framework for Safe Reinforcement Learning with Markov Games of Intervention

David Henry Mguni,Joel Jennings,Taher Jafferjee,Aivar Sootla,Changmin Yu,Usman Islam,Ziyan Wang,Yaodong Yang,Jun Wang

Exploring in an unknown system can place an agent in dangerous situations, exposing to potentially catastrophic hazards. Many current approaches for tackling safe learning in reinforcement learning (RL) lead to a trade-off between safe exploration and fulfilling the task. Though these methods possibly incur fewer safety violations they often also lead to reduced task performance. In this paper, we take the first step in introducing a generation of RL solvers that learn to minimise safety violations while maximising the task reward to the extend that can be tolerated by safe policies. Our approach uses a new two-player framework for safe RL called DESTA. The core of DESTA is a novel game between two RL agents: Safety Agent that is delegated the task of minimising safety violations and Task Agent whose goal is to maximise the reward set by the environment task. Safety Agent can selectively take control of the system at any given point to prevent safety violations while Task Agent is free to execute its actions at all other states. This framework enables Safety Agent to learn to take actions that minimise future safety violations (during and after training) by performing safe actions at certain states while Task Agent performs actions that maximise the task performance everywhere else. We demonstrate DESTA’s ability to tackle challenging tasks and compare against state-of-the-art RL methods in Safety Gym Benchmarks which simulate real-world physical systems and OpenAI’s Lunar Lander.

****

LIGS: Learnable Intrinsic-Reward Generation Selection for Multi-Agent Learning

David Henry Mguni,Taher Jafferjee,Jianhong Wang,Nicolas Perez-Nieves,Oliver Slumbers,Feifei Tong,Yang Li,Jiangcheng Zhu,Yaodong Yang,Jun Wang

Efficient exploration is important for reinforcement learners (RL) to achieve high rewards. In multi-agent systems, coordinated exploration and behaviour is critical for agents to jointly achieve optimal outcomes. In this paper, we introduce a new general framework for improving coordination and performance of multi-agent reinforcement learners (MARL). Our framework, named Learnable Intrinsic-Reward Generation Selection algorithm (LIGS) introduces an adaptive learner, Generator that observes the agents and learns to construct intrinsic rewards online that coordinate the agents’ joint exploration and joint behaviour. Using a novel combination of reinforcement learning (RL) and switching controls, LIGS determines the best states to learn to add intrinsic rewards which leads to a highly efficient learning process. LIGS can subdivide complex tasks making them easier to solve and enables systems of RL agents to quickly solve environments with sparse rewards. LIGS can seamlessly adopt existing multi-agent RL algorithms and our theory shows that it ensures convergence to joint policies that deliver higher system performance. We demonstrate the superior performance of the LIGS framework in challenging tasks in Foraging and StarCraft II and show LIGS is capable of tackling tasks previously unsolvable by MARL methods.

****

Bayesian Modeling and Uncertainty Quantification for Learning to Optimize: What, Why, and How

Yuning You,Yue Cao,Tianlong Chen,Zhangyang Wang,Yang Shen

Optimizing an objective function with uncertainty awareness is well-known to improve the accuracy and confidence of optimization solutions. Meanwhile, another relevant but very different question remains yet open: how to model and quantify the uncertainty of an optimization algorithm (a.k.a., optimizer) itself? To close such a gap, the prerequisite is to consider the optimizers as sampled from a distribution, rather than a few prefabricated and fixed update rules. We first take the novel angle to consider the algorithmic space of optimizers, and provide definitions for the optimizer prior and likelihood, that intrinsically determine the posterior and therefore uncertainty. We then leverage the recent advance of learning to optimize (L2O) for the space parameterization, with the end-to-end training pipeline built via variational inference, referred to as uncertainty-aware L2O (UA-L2O). Our study represents the first effort to recognize and quantify the uncertainty of the optimization algorithm. The extensive numerical results show that, UA-L2O achieves superior uncertainty calibration with accurate confidence estimation and tight confidence intervals, suggesting the improved posterior estimation thanks to considering optimizer uncertainty. Intriguingly, UA-L2O even improves optimization performances for two out of three test functions, the loss function in data privacy attack, and four of five cases of the energy function in protein docking. Our codes are released at https://github.com/Shen-Lab/Bayesian-L2O.

****

On the relation between statistical learning and perceptual distances

Alexander Hepburn,Valero Laparra,Raul Santos-Rodriguez,Johannes Ballé,Jesus Malo

It has been demonstrated many times that the behavior of the human visual system is connected to the statistics of natural images. Since machine learning relies on the statistics of training data as well, the above connection has interesting implications when using perceptual distances (which mimic the behavior of the human visual system) as a loss function. In this paper, we aim to unravel the non-trivial relationships between the probability distribution of the data, perceptual distances, and unsupervised machine learning. To this end, we show that perceptual sensitivity is correlated with the probability of an image in its close neighborhood. We also explore the relation between distances induced by autoencoders and the probability distribution of the training data, as well as how these induced distances are correlated with human perception. Finally, we find perceptual distances do not always lead to noticeable gains in performance over Euclidean distance in common image processing tasks, except when data is scarce and the perceptual distance provides regularization. We propose this may be due to a double-counting effect of the image statistics, once in the perceptual distance and once in the training procedure.

****

On the interventional consistency of autoencoders

Giulia Lanzillotta,Felix Leeb,Stefan Bauer,Bernhard Schölkopf

Autoencoders have played a crucial role in the field of representation learning since its inception, proving to be a flexible learning scheme able to accommodate various notions of optimality of the representation. The now established idea of disentanglement and the recently popular perspective of causality in representation learning identify modularity and robustness to be essential characteristics of the optimal representation. In this work, we show that the current conceptual tools available to assess the quality of the representation against these criteria (e.g. latent traversals or disentanglement metrics) are inadequate. In this regard, we introduce the notion of \emph{interventional consistency} of a representation and argue that it is a desirable property of any disentangled representation. We develop a general training scheme for autoencoders that takes into account interventional consistency in the optimality condition. We present empirical evidence toward the validity of the approach on three different autoencoders, namely standard autoencoders (AE), variational autoencoders (VAE) and structural autoencoders (SAE).

Another key finding in this work is that differentiating between information and structure in the latent space of autoencoders can increase the modularity and i nterpretability of the resulting representation.

****

Deep Ensembling with No Overhead for either Training or Testing: The All-Round Blessings of Dynamic Sparsity

Shiwei Liu,Tianlong Chen,Zahra Atashgahi,Xiaohan Chen,Ghada Sokar,Elena Mocanu,Mykola Pechenizkiy,Zhangyang Wang,Decebal Constantin Mocanu

The success of deep ensembles on improving predictive performance, uncertainty estimation, and out-of-distribution robustness has been extensively studied in the machine learning literature. Albeit the promising results, naively training multiple deep neural networks and combining their predictions at inference leads to prohibitive computational costs and memory requirements. Recently proposed efficient ensemble approaches reach the performance of the traditional deep ensembles with significantly lower costs. However, the training resources required by these approaches are still at least the same as training a single dense model. In this work, we draw a unique connection between sparse neural network training and deep ensembles, yielding a novel efficient ensemble learning framework called $FreeTickets$. Instead of training multiple dense networks and averaging them, we directly train sparse subnetworks from scratch and extract diverse yet accurate subnetworks during this efficient, sparse-to-sparse training. Our framework, $FreeTickets$, is defined as the ensemble of these relatively cheap sparse subnetworks. Despite being an ensemble method, $FreeTickets$ has even fewer parameters and training FLOPs than a single dense model. This seemingly counter-intuitive outcome is due to the ultra training/inference efficiency of dynamic sparse training. $FreeTickets$ surpasses the dense baseline in all the following criteria: prediction accuracy, uncertainty estimation, out-of-distribution (OoD) robustness, as well as efficiency for both training and inference. Impressively, $FreeTickets$ outperforms the naive deep ensemble with ResNet50 on ImageNet using around only $1/5$ of the training FLOPs required by the latter. We have released our source code at https://github.com/VITA-Group/FreeTickets.

****

Logarithmic Unbiased Quantization: Practical 4-bit Training in Deep Learning

Brian Chmiel,Ron Banner,Elad Hoffer,Hilla Ben Yaacov,Daniel Soudry

Quantization of the weights and activations is one of the main methods to reduce the computational footprint of Deep Neural Networks (DNNs) training. Current methods enable 4-bit quantization of the forward phase. However, this constitutes only a third of the training process. Reducing the computational footprint of the entire training process requires the quantization of the neural gradients, i.e., the loss gradients with respect to the outputs of intermediate neural layers.

In this work, we examine the importance of having unbiased quantization in quantized neural network training, where to maintain it, and how. Based on this, we suggest a logarithmic unbiased quantization (LUQ) method to quantize both the forward and backward phase to 4-bit, achieving state-of-the-art results in 4-bit training. For example, in ResNet50 on ImageNet, we achieved a degradation of 1.18 %; we further improve this to degradation of only 0.64 % after a single epoch of high precision fine-tuning combined with a variance reduction method. Finally, we suggest a method that exploits the low precision format by avoiding multiplications during two-thirds of the training process, thus reducing by 5x the area used by the multiplier. A reference implementation is supplied in the supplementary material.

****

Towards Federated Learning on Time-Evolving Heterogeneous Data

Yongxin Guo,Tao Lin,Xiaoying Tang

Federated Learning (FL) is an emerging learning paradigm that preserves privacy by ensuring client data locality on edge devices. The optimization of FL is challenging in practice due to the diversity and heterogeneity of the learning system. Despite recent research efforts on improving the optimization of heterogeneous data, the impact of time-evolving heterogeneous data in real-world scenarios, such as changing client data or intermittent clients joining or leaving during training, has not been well studied.

In this work, we propose Continual Federated Learning (CFL), a flexible framework, to capture the time-evolving heterogeneity of FL. CFL covers complex and realistic scenarios---which are challenging to evaluate in previous FL formulations---by extracting the information of past local datasets and approximating the local objective functions. Theoretically, we demonstrate that CFL methods achieve a faster convergence rate than FedAvg in time-evolving scenarios, with the benefit being dependent on approximation quality. In a series of experiments, we show that the numerical findings match the convergence analysis, and CFL methods significantly outperform the other SOTA FL baselines.

****

Distributed Optimal Margin Distribution Machine

Yilin Wang,nan cao,Teng Zhang,Hai Jin

Optimal margin Distribution Machine (ODM), a newly proposed statistical learning framework rooting in the novel margin theory, demonstrates better generalization performance than the traditional large margin based counterparts. Nonetheless, the same with other kernel methods, it suffers from the ubiquitous scalability problem in terms of both computation time and memory. In this paper, we propose a Distributed solver for ODM (DiODM), which leads to nearly ten times speedup for training kernel ODM. It exploits a novel data partition method to make the local ODM trained on each partition has a solution close to the global one. When linear kernel used, we extend a communication efficient distributed SVRG method to further accelerate the training. Extensive empirical studies validate the superiority of our proposed method compared to other off-the-shelf distributed quadratic programming solvers for kernel methods.

****

BWCP: Probabilistic Learning-to-Prune Channels for ConvNets via Batch Whitening

Wenqi Shao,Hang Yu,Zhaoyang Zhang,Hang Xu,Zhenguo Li,Ping Luo

This work presents a probabilistic channel pruning method to accelerate Convolutional Neural Networks (CNNs). Previous pruning methods often zero out unimportant channels in training in a deterministic manner, which reduces CNN's learning capacity and results in suboptimal performance. To address this problem, we develop a probability-based pruning algorithm, called batch whitening channel pruning (BWCP), which can stochastically discard unimportant channels by modeling the probability of a channel being activated. BWCP has several merits. (1) It simultaneously trains and prunes CNNs from scratch in a probabilistic way, exploring larger network space than deterministic methods. (2) BWCP is empowered by the proposed batch whitening tool, which is able to empirically and theoretically increase the activation probability of useful channels while reducing the probability of unimportant channels without adding any extra parameters and computational cost in inference. (3) Extensive experiments on CIFAR-10, CIFAR-100, and ImageNet with various network architectures show that BWCP outperforms its counterparts by achieving better accuracy given limited computational budgets. For example, ResNet50 pruned by BWCP has only 0.58% Top-1 accuracy drop on ImageNet, while reducing 42.9% FLOPs of the plain ResNet50.

****

Deep Inverse Reinforcement Learning via Adversarial One-Class Classification

Daiko Kishikawa,Sachiyo Arai

Traditional inverse reinforcement learning (IRL) methods require a loop to find the optimal policy for each reward update (called an inner loop), resulting in very time-consuming reward estimation. In contrast, classification-based IRL methods, which have been studied recently, do not require an inner loop and estimate rewards quickly, although it is difficult to prepare an appropriate baseline corresponding to the expert trajectory. In this study, we introduced adversarial one-class classification into the classification-based IRL framework, and consequently developed a novel IRL method that requires only expert trajectories. We experimentally verified that the developed method can achieve the same performance as existing methods.

****

Denoising Diffusion Gamma Models

Eliya Nachmani,Robin San Roman,Lior Wolf

Generative diffusion processes are an emerging and effective tool for image and speech generation. In the existing methods, the underlying noise distribution of the diffusion process is Gaussian noise. However, fitting distributions with more degrees of freedom could improve the performance of such generative models. In this work, we investigate other types of noise distribution for the diffusion process. Specifically, we introduce the Denoising Diffusion Gamma Model (DDGM) and show that noise from Gamma distribution provides improved results for image and speech generation. Our approach preserves the ability to efficiently sample state in the training diffusion process while using Gamma noise.

****

BoolNet: Streamlining Binary Neural Networks Using Binary Feature Maps

Nianhui Guo,Joseph Bethge,Haojin Yang,Kai Zhong,Xuefei Ning,Christoph Meinel,Yu Wang

Recent works on Binary Neural Networks (BNNs) have made promising progress in narrowing the accuracy gap of BNNs to their 32-bit counterparts, often based on specialized model designs using additional 32-bit components. Furthermore, most previous BNNs use 32-bit values for feature maps and residual shortcuts, which helps to maintain the accuracy, but is not friendly to hardware accelerators with limited memory, energy, and computing resources. Thus, we raise the following question: How can accuracy and energy consumption be balanced in a BNN design? We extensively study this fundamental problem in this work and propose BoolNet: an architecture without most commonly used 32-bit components that uses 1-bit values to store feature maps. Experimental results on ImageNet demonstrate that BoolNet can achieve 63.0% Top-1 accuracy coupled with an energy reduction of 2.95x compared to recent state-of-the-art BNN architectures. Code and trained models are available at: (URL in the final version).

****

Implicit Bias of Projected Subgradient Method Gives Provable Robust Recovery of Subspaces of Unknown Codimension

Paris Giampouras,Benjamin David Haeffele,Rene Vidal

Robust subspace recovery (RSR) is the problem of learning a subspace from sample data points corrupted by outliers. Dual Principal Component Pursuit (DPCP) is a robust subspace recovery method that aims to find a basis for the orthogonal complement of the subspace by minimizing the sum of the distances of the points to the subspaces subject to orthogonality constraints on the basis. Prior work has shown that DPCP can provably recover the correct subspace in the presence of outliers as long as the true dimension of the subspace is known. In this paper, we show that if the orthogonality constraints --adopted in previous DPCP formulations-- are relaxed and random initialization is used instead of spectral one, DPCP can provably recover a subspace of \emph{unknown dimension}. Specifically, we propose a very simple algorithm based on running multiple instances of a projected sub-gradient descent method (PSGM), with each problem instance seeking to find one vector in the null space of the subspace. We theoretically prove that under mild conditions this approach succeeds with high probability. In particular, we show that 1) all of the problem instances will converge to a vector in the nullspace of the subspace and 2) the ensemble of problem instance solutions will be sufficiently diverse to fully span the nullspace of the subspace thus also revealing its true unknown codimension. We provide empirical results that corroborate our theoretical results and showcase the remarkable implicit rank regularization behavior of the PSGM algorithm that allows us to perform RSR without knowing the subspace dimension

****

HyperDQN: A Randomized Exploration Method for Deep Reinforcement Learning

Ziniu Li,Yingru Li,Yushun Zhang,Tong Zhang,Zhi-Quan Luo

Randomized least-square value iteration (RLSVI) is a provably efficient exploration method. However, it is limited to the case where (1) a good feature is known in advance and (2) this feature is fixed during the training. If otherwise, RLSVI suffers an unbearable computational burden to obtain the posterior samples. In this work, we present a practical algorithm named HyperDQN to address the above issues under deep RL. In addition to a non-linear neural network (i.e., base m odel) that predicts Q-values, our method employs a probabilistic hypermodel (i.e., meta model), which outputs the parameter of the base model. When both models are jointly optimized under a specifically designed objective, three purposes can be achieved. First, the hypermodel can generate approximate posterior samples regarding the parameter of the Q-value function. As a result, diverse Q-value functions are sampled to select exploratory action sequences. This retains the punchline of RLSVI for efficient exploration. Second, a good feature is learned to approximate Q-value functions. This addresses limitation (1). Third, the posterior samples of the Q-value function can be obtained in a more efficient way than the existing methods, and the changing feature does not affect the efficiency. This deals with limitation (2). On the Atari suite, HyperDQN with 20M frames outperforms DQN with 200M frames in terms of the maximum human-normalized score. For SuperMarioBros, HyperDQN outperforms several exploration bonus and randomized exploration methods on 5 out of 9 games.

****

Unraveling Model-Agnostic Meta-Learning via The Adaptation Learning Rate

Yingtian Zou,Fusheng Liu,Qianxiao Li

Model-Agnostic Meta-Learning (MAML) aims to find initial weights that allow fast adaptation to new tasks. The adaptation (inner loop) learning rate in MAML plays a central role in enabling such fast adaptation. However, how to choose this value in practice and how this choice affects the adaptation error remains less explored. In this paper, we study the effect of the adaptation learning rate in meta-learning with mixed linear regression. First, we present a principled way to estimate optimal adaptation learning rates that minimize the population risk of MAML. Second, we interpret the underlying dependence between the optimal adaptation learning rate and the input data. Finally, we prove that compared with empirical risk minimization (ERM), MAML produces an initialization with a smaller average distance to the task optima, consistent with previous practical findings. These results are corroborated with numerical experiments.

****

Graph Tree Neural Networks

Seokjun Kim,Jaeeun Jang,Heeseok Jung,Hyeoncheol Kim

In the field of deep learning, various architectures have been developed. However, most studies are limited to specific tasks or datasets due to their fixed layer structure. In this paper, we do not express the structure delivering information as a network model but as a data structure called a graph tree. And we propose two association models of graph tree neural networks(GTNNs) designed to solve the problems of existing networks by analyzing the structure of human neural networks. Defining the starting and ending points in a single graph is difficult, and a tree cannot express the relationship among sibling nodes. On the contrary, a graph tree(GT) can express leaf and root nodes as its starting and ending points and the relationship among sibling nodes. Instead of using fixed sequence layers, we create a GT for each data and train GTNN according to the tree's structure. GTNNs are data-driven learning in which the number of convolutions varies according to the depth of the tree. Moreover, these models can simultaneously learn various types of datasets through the recursive learning method. Depth-first convolution (DFC) encodes the interaction result from leaf nodes to the root node in a bottom-up approach, and depth-first deconvolution (DFD) decodes the interaction result from the root node to the leaf nodes in a top-down approach. To demonstrate the performance of these networks, we conducted two experiments. The first experiment is whether various datasets can be processed by combining GTNN and feature extraction networks(processing various datasets). The second experiment is about whether the output of GTNN can embed information on all data contained in the GT(association). We compared the performance of existing networks that separately learned image, sound, and natural language datasets with the performance simultaneously learned by connecting these networks. As a result, these models learned without significant performance degradation, and the output vector contained all the information in the GT.

****

Information-Theoretic Generalization Bounds for Iterative Semi-Supervised Learni ng

Haiyun He,Hanshu YAN,Vincent Tan

We consider iterative semi-supervised learning (SSL) algorithms that iteratively generate pseudo-labels for a large amount unlabelled data to progressively refine the model parameters. In particular, we seek to understand the behaviour of the {\em generalization error} of iterative SSL algorithms using information-theoretic principles. To obtain bounds that are amenable to numerical evaluation, we first work with a simple model---namely, the binary Gaussian mixture model. Our theoretical results suggest that when the class conditional variances are not too large, the upper bound on the generalization error decreases monotonically with the number of iterations, but quickly saturates. The theoretical results on the simple model are corroborated by extensive experiments on several benchmark datasets such as the MNIST and CIFAR datasets in which we notice that the generalization error improves after several pseudo-labelling iterations, but saturates afterwards.

****

Agnostic Personalized Federated Learning with Kernel Factorization

Wonyong Jeong,Sung Ju Hwang

Considering the futuristic scenarios of federated learning at a worldwide scale, it is highly probable that local participants can have their own personalized labels, which might not be compatible with each other even for the same class, and can be also possibly from a variety of multiple domains. Nevertheless, they should be benefited from others while selectively taking helpful knowledge. Toward such extreme scenarios of federated learning, however, most existing approaches are limited in that they often assume: (1) labeling schemes are all synchronized amongst clients; (2) the local data is from the same single dataset (domain). In this sense, we introduce an intensively realistic problem of federated learning, namely Agnostic Personalized Federated Learning (APFL), where any clients, regardless of what they have learned with their personalized labels, can collaboratively learn while benefiting each other. We then study two essential challenges of the agnostic personalized federated learning, which are (1) Label Heterogeneity where local clients learn from the same single domain but labeling schemes are not synchronized with each other and (2) Domain Heterogeneity where the clients learn from the different datasets which can be semantically similar or dissimilar for each other. To tackle these problems, we propose our novel method, namely Similarity Matching and Kernel Factorization (SimFed). Our method measures semantic similarity/dissimilarity between locally learned knowledge and matches/aggregates the relevant ones that are beneficial to each other. Furthermore, we factorize our model parameters into two basis vectors and the sparse masks to effectively capture permutation-robust representations and reduce information loss when aggregating the heterogeneous knowledge. We exhaustively validate our method on both single- and multi-domain datasets, showing that our method outperforms the current state-of-the-art federated learning methods.

****

On the Optimal Memorization Power of ReLU Neural Networks

Gal Vardi,Gilad Yehudai,Ohad Shamir

We study the memorization power of feedforward ReLU neural networks. We show that such networks can memorize any N points that satisfy a mild separability assumption using $\tilde{O}\left(\sqrt{N}\right)$ parameters. Known VC-dimension upper bounds imply that memorizing N samples requires $\Omega(\sqrt{N})$ parameters, and hence our construction is optimal up to logarithmic factors. We also give a generalized construction for networks with depth bounded by $1 \leq L \leq \sqrt{N}$, for memorizing N samples using $\tilde{O}(N/L)$ parameters. This bound is also optimal up to logarithmic factors. Our construction uses weights with large bit complexity. We prove that having such a large bit complexity is both necessary and sufficient for memorization with a sub-linear number of parameters.

****

Task Conditioned Stochastic Subsampling

Bruno Andreis,Seanie Lee,A. Tuan Nguyen,Juho Lee,Eunho Yang,Sung Ju Hwang

Deep Learning algorithms are designed to operate on huge volumes of high dimensional data such as images. In order to reduce the volume of data these algorithms must process, we propose a set-based two-stage end-to-end neural subsampling model that is jointly optimized with an \textit{arbitrary} downstream task network such as a classifier. In the first stage, we efficiently subsample \textit{candidate elements} using conditionally independent Bernoulli random variables, followed by conditionally dependent autoregressive subsampling of the candidate elements using Categorical random variables in the second stage. We apply our method to feature and instance selection and show that our method outperforms the relevant baselines under very low subsampling rates on many tasks including image classification, image reconstruction, function reconstruction and few-shot classification. Additionally, for nonparametric models such as Neural Processes that require to leverage whole training data at inference time, we show that our method enhances the scalability of these models. To ensure easy reproducibility, we provide source code in the \textbf{Supplementary Material}.

****

ST-DDPM: Explore Class Clustering for Conditional Diffusion Probabilistic Models
Zhijie Lin,Zijian Zhang,Zhou Zhao
Score-based generative models involve sequentially corrupting the data distribution with noise and then learns to recover the data distribution based on score matching. In this paper, for the diffusion probabilistic models, we first delve into the changes of data distribution during the forward process of the Markov chain and explore the class clustering phenomenon. Inspired by the class clustering phenomenon, we devise a novel conditional diffusion probabilistic model by explicitly modeling the class center in the forward and reverse process, and make an elegant modification to the original formulation, which enables controllable generation and gets interpretability. We also provide another direction for faster sampling and more analysis of our method. To verify the effectiveness of the formulated framework, we conduct extensive experiments on multiple tasks, and achieve competitive results compared with the state-of-the-art methods(conditional image generation on CIFAR-10 with an inception score of 9.58 and FID score of 3.05).

****

Learned Index with Dynamic ϵ
Daoyuan Chen,Wuchao Li,Yaliang Li,Bolin Ding,Kai Zeng,Defu Lian,Jingren Zhou
Index structure is a fundamental component in database and facilitates broad data retrieval applications. Recent learned index methods show superior performance by learning hidden yet useful data distribution with the help of machine learning, and provide a guarantee that the prediction error is no more than a pre-defined ϵ. However, existing learned index methods adopt a fixed ϵ for all the learned segments, neglecting the diverse characteristics of different data localities. In this paper, we propose a mathematically-grounded learned index framework with dynamic ϵ, which is efficient and pluggable to existing learned index methods. We theoretically analyze prediction error bounds that link ϵ with data characteristics for an illustrative learned index method. Under the guidance of the derived bounds, we learn how to vary ϵ and improve the index performance with a better space-time trade-off. Experiments with real-world datasets and several state-of-the-art methods demonstrate the efficiency, effectiveness and usability of the proposed framework.

****

A Deep Latent Space Model for Directed Graph Representation Learning
Hanxuan Yang,Qingchao Kong,Wenji Mao
Graph representation learning is a fundamental problem for modeling relational data and benefits a number of downstream applications. Traditional Bayesian-based random graph models and recent deep learning based methods are complementary to each other in interpretability and scalability. To take the advantages of both models, some combined methods have been proposed. However, existing models are mainly designed for \textit{undirected graphs}, while a large portion of real-world graphs are directed. The focus of this paper is on \textit{directed graphs}.

We propose a Deep Latent Space Model (DLSM) for directed graphs to incorporate the traditional latent space random graph model into deep learning frameworks via a hierarchical variational auto-encoder architecture. To adapt to directed graphs, our model generates multiple highly interpretable latent variables as node representations, and the interpretability of representing node influences is theoretically proved. Moreover, our model achieves good scalability for large graphs via the fast stochastic gradient variational Bayes inference algorithm. The experimental results on real-world graphs demonstrate that our proposed model achieves the state-of-the-art performances on link prediction and community detection tasks while generating interpretable node representations.

****

Self-consistent Gradient-like Eigen Decomposition in Solving Schrödinger Equations

Xihan Li,Xiang Chen,Rasul Tutunov,Haitham Bou Ammar,Lei Wang,Jun Wang

The Schrödinger equation is at the heart of modern quantum mechanics. Since exact solutions of the ground state are typically intractable, standard approaches approximate Schrödinger's equation as forms of nonlinear generalized eigenvalue problems $F(V)V = SV\Lambda$ in which $F(V)$, the matrix to be decomposed, is a function of its own top-k smallest eigenvectors V, leading to a ``self-consistency problem''. Traditional iterative methods heavily rely on high-quality initial guesses of V generated via domain-specific heuristics methods based on quantum mechanics. In this work, we eliminate such a need for domain-specific heuristics by presenting a novel framework, Self-consistent Gradient-like Eigen Decomposition (SCGLED) that regards $F(V)$ as a special ``online data generator'', thus allows gradient-like eigendecomposition methods in streaming k-PCA to approach the self-consistency of the equation from scratch in an iterative way similar to online learning. With several critical numerical improvements, SCGLED is robust to initial guesses, free of quantum-mechanism-based heuristics designs, and neat in implementation. Our experiments show that it not only can simply replace traditional heuristics-based initial guess methods with large performance advantage (achieved averagely 25x more precise than the best baseline in similar wall time), but also is capable of finding highly precise solutions independently without any traditional iterative methods.

****

AARL: Automated Auxiliary Loss for Reinforcement Learning

Tairan He,Yuge Zhang,Kan Ren,Che Wang,Weinan Zhang,Dongsheng Li,Yuqing Yang

A good state representation is crucial to reinforcement learning (RL) while an ideal representation is hard to learn only with signals from the RL objective. Thus, many recent works manually design auxiliary losses to improve sample efficiency and decision performance. However, handcrafted auxiliary losses rely heavily on expert knowledge, and therefore lack scalability and can be suboptimal for boosting RL performance. In this work, we introduce Automated Auxiliary loss for Reinforcement Learning (AARL), a principled approach that automatically searches the optimal auxiliary loss function for RL. Specifically, based on the collected trajectory data, we define a general auxiliary loss space of size 4.6×10^{19} and explore the space with an efficient evolutionary search strategy. We evaluate AARL on the DeepMind Control Suite and show that the searched auxiliary losses have significantly improved RL performance in both pixel-based and state-based settings, with the largest performance gain observed in the most challenging tasks. AARL greatly outperforms state-of-the-art methods and demonstrates strong generalization ability in unseen domains and tasks. We further conduct extensive studies to shed light on the effectiveness of auxiliary losses in RL.

****

iFlood: A Stable and Effective Regularizer

Yuexiang Xie,Zhen WANG,Yaliang Li,Ce Zhang,Jingren Zhou,Bolin Ding

Various regularization methods have been designed to prevent overfitting of machine learning models. Among them, a surprisingly simple yet effective one, called Flooding, is proposed recently, which directly constrains the training loss on average to stay at a given level. However, our further studies uncover that the design of the loss function of Flooding can lead to a discrepancy between its ob jective and implementation, and cause the instability issue. To resolve these issues, in this paper, we propose a new regularizer, called individual Flood (denoted as iFlood). With instance-level constraints on training loss, iFlood encourages the trained models to better fit the under-fitted instances while suppressing the confidence on over-fitted ones. We theoretically show that the design of iFlood can be intrinsically connected with removing the noise or bias in training data, which makes it suitable for a variety of applications to improve the generalization performances of learned models. We also theoretically link iFlood to some other regularizers by comparing the inductive biases they introduce. Our experimental results on both image classification and language understanding tasks confirm that models learned with iFlood can stably converge to solutions with better generalization ability, and behave consistently at instance-level.

****

Hypergraph Convolutional Networks via Equivalency between Hypergraphs and Undirected Graphs

Jiying Zhang,Fuyang Li,Xi Xiao,Tingyang Xu,Yu Rong,Junzhou Huang,Yatao Bian

As a powerful tool for modeling the complex relationships, hypergraphs are gaining popularity from the graph learning community. However, commonly used algorithms in deep hypergraph learning were not specifically designed for hypergraphs with edge-dependent vertex weights (EDVWs). To fill this gap, we build the equivalency condition between EDVW-hypergraphs and undirected simple graphs, which enables utilizing existing undirected graph neural networks as subroutines to learn high-order interactions induced by EDVWs of hypergraphs. Specifically, we define a generalized hypergraph with vertex weights by proposing a unified random walk framework, under which we present the equivalency condition between generalized hypergraphs and undigraphs. Guided by the equivalency results, we propose a Generalized Hypergraph Convolutional Network (GHCN) architecture for deep hypergraph learning. Furthermore, to improve the long-range interactions and alleviate the over-smoothing issue, we further propose the Simple Hypergraph Spectral Convolution (SHSC) model by constructing the Discounted Markov Diffusion Kernel from our random walk framework. Extensive experiments from various domains including social network analysis, visual objective classification, and protein fold classification demonstrate that the proposed approaches outperform state-of-the-art spectral methods with a large margin.

****

FlexConv: Continuous Kernel Convolutions With Differentiable Kernel Sizes

David W. Romero,Robert-Jan Bruintjes,Jakub Mikolaj Tomczak,Erik J Bekkers,Mark Hoogendoorn,Jan van Gemert

When designing Convolutional Neural Networks (CNNs), one must select the size of the convolutional kernels before training. Recent works show CNNs benefit from different kernel sizes at different layers, but exploring all possible combinations is unfeasible in practice. A more efficient approach is to learn the kernel size during training. However, existing works that learn the kernel size have a limited bandwidth. These approaches scale kernels by dilation, and thus the detail they can describe is limited. In this work, we propose FlexConv, a novel convolutional operation with which high bandwidth convolutional kernels of learnable kernel size can be learned at a fixed parameter cost. FlexNets model long-term dependencies without the use of pooling, achieve state-of-the-art performance on several sequential datasets, outperform recent works with learned kernel sizes, and are competitive with much deeper ResNets on image benchmark datasets. Additionally, FlexNets can be deployed at higher resolutions than those seen during training. To avoid aliasing, we propose a novel kernel parameterization with which the frequency of the kernels can be analytically controlled. Our novel kernel parameterization shows higher descriptive power and faster convergence speed than existing parameterizations. This leads to important improvements in classification accuracy.

****

OUMG: Objective and Universal Metric for Text Generation with Guiding Ability

Hanxu Liu,Nianmin Yao

Existing evaluation metrics for text generation rely on comparing candidate sent ences to reference sentences. Some text generation tasks, such as story generation and poetry generation, have no fixed optimal answer and cannot match a corresponding reference for each sentence. Therefore, there is a lack of an objective and universal evaluation metric. To this end, we propose OUMG, a general metric that does not depend on reference standards. We train a discriminator to distinguish between human-generated and machine-generated text, which is used to score the sentences generated by the model. These scores reflect how similar the sentences are to human-generated texts. The capability of the discriminator can be measured by its accuracy, so it avoids the subjectivity of human judgments. Furthermore, the trained discriminator can also guide the text generation process to improve model performance. Experiments on poetry generation demonstrate that OUMG can objectively evaluate text generation models without reference standards. After combining the discriminator with the generation model, the original model can produce significantly higher quality results.

****

Edge Rewiring Goes Neural: Boosting Network Resilience via Policy Gradient

Shanchao Yang,MA KAILI,Baoxiang Wang,Hongyuan Zha

Improving the resilience of a network protects the system from natural disasters and malicious attacks.

This is typically achieved by introducing new edges, which however may reach beyond the maximum number of connections a node could sustain.

Many studies then resort to the degree-preserving operation of rewiring, which swaps existing edges AC, BD to new edges AB, CD.

A significant line of studies focuses on this technique for theoretical and practical results while leaving three limitations: network utility loss, local optimality, and transductivity.

In this paper, we propose ResiNet, a reinforcement learning (RL)-based framework to discover Resilient Network topologies against various disasters and attacks.

ResiNet is objective agnostic which allows the utility to be balanced by incorporating it into the objective function.

The local optimality, typically seen in greedy algorithms, is addressed by casting the cumulative resilience gain into a sequential decision process of step-wise rewiring.

The transductivity, which refers to the necessity to run a computationally intensive optimization for each input graph, is lifted by our variant of RL with auto-regressive permutation-invariant variable action space.

ResiNet is armed by our technical innovation, Filtration enhanced GNN (FireGNN), which distinguishes graphs with minor differences.

It is thus possible for ResiNet to capture local structure changes and adapt its decision among consecutive graphs, which is known to be infeasible for GNN.

Extensive experiments demonstrate that with a small number of rewiring operations, ResiNet achieves a near-optimal resilience gain on multiple graphs while balancing the utility, with a large margin compared to existing approaches.

****

A Comprehensive Overhaul of Distilling Unconditional GANs

Guodong Xu,Yuenan Hou,Ziwei Liu,Chen Change Loy

Generative adversarial networks (GANs) have achieved impressive results on various content generation tasks. Yet, their high demand on storage and computation impedes their deployment on resource-constrained devices. Though several GAN compression methods have been proposed to address the problem, most of them focus on conditional GANs. In this paper, we provide a comprehensive overhaul of distilling unconditional GAN, especially for the popular StyleGAN2 architecture. Our key insight is that the main challenge of unconditional GAN distillation lies in the output discrepancy issue, where the teacher and student model yield different outputs given the same input latent code. Standard knowledge distillation losses typically fail under this heterogeneous distillation scenario. We conduct thorough analysis about the reasons and effects of this discrepancy issue, and identify that the style module plays a vital role in determining semantic information of generated images. Based on this finding, we propose a novel initialization s

trategy for the student model, which can ensure the output consistency to the maximum extent. To further enhance the semantic consistency between the teacher and student model, we present another latent-direction-based distillation loss that preserves the semantic relations in latent space. Extensive experiments demonstrate that our framework achieves state-of-the-art results in StyleGAN2 distillation, outperforming the existing GAN distillation methods by a large margin.
****
Safe Linear-Quadratic Dual Control with Almost Sure Performance Guarantee
Yiwen Lu,Yilin Mo
This paper considers the linear-quadratic dual control problem where the system parameters need to be identified and the control objective needs to be optimized in the meantime. Contrary to existing works on data-driven linear-quadratic regulation, which typically provide error or regret bounds within a certain probability, we propose an online algorithm that guarantees the asymptotic optimality of the controller in the almost sure sense. Our dual control strategy consists of two parts: a switched controller with time-decaying exploration noise and Markov parameter inference based on the cross-correlation between the exploration noise and system output. Central to the almost sure performance guarantee is a safe switched control strategy that falls back to a known conservative but stable controller when the actual state deviates significantly from the target state. We prove that this switching strategy rules out any potential destabilizing controllers from being applied, while the performance gap between our switching strategy and the optimal linear state feedback is exponentially small. Under our dual control scheme, the parameter inference error scales as $O(T^{-1/4+\epsilon})$, while the suboptimality gap of control performance scales as $O(T^{-1/2+\epsilon})$, where T is the number of time steps, and ϵ is an arbitrarily small positive number. Simulation results on an industrial process example are provided to illustrate the effectiveness of our proposed strategy.
****
Density-based Clustering with Kernel Diffusion
Chao Zheng,Yingjie Chen,Chong Chen,Jianqiang Huang,Xian-Sheng Hua
Finding a suitable density function is essential for density-based clustering algorithms such as DBSCAN and DPC. A naive density corresponding to the indicator function of a unit d-dimensional Euclidean ball is commonly used in these algorithms. Such density suffers from capturing local features in complex datasets. To tackle this issue, we propose a new kernel diffusion density function, which is adaptive to data of varying local distributional characteristics and smoothness. Furthermore, we develop a surrogate that can be efficiently computed in linear time and space and prove that it is asymptotically equivalent to the kernel diffusion density function. Extensive empirical experiments on benchmark and large-scale face image datasets show that the proposed approach not only achieves a significant improvement over classic density-based clustering algorithms but also outperforms the state-of-the-art face clustering methods by a large margin.
****
Spatially Invariant Unsupervised 3D Object-Centric Learning and Scene Decomposition
Tianyu Wang,miaomiao Liu,Kee Siong Ng
We tackle the problem of deep object-centric learning from a point cloud which is crucial for high-level relational reasoning and scalable machine intelligence.

In particular, we introduce a framework, SPAIR3D, to factorize a 3D point cloud into a spatial mixture model where each component corresponds to one object.
To model the spatial mixture model on point clouds, we derive the Chamfer Mixture Loss, which fits naturally into our variational training pipeline. Moreover, we adopt an object-specification scheme that describes each object’s location relative to its local voxel grid cell.
Such a scheme allows SPAIR3D to model scenes with an arbitrary number of objects.
We evaluate our method on the task of unsupervised scene decomposition.
Experimental results demonstrate that SPAIR3D has strong scalability and is capa

ble of detecting and segmenting an unknown number of objects from a point cloud in an unsupervised manner.
****

Zero Pixel Directional Boundary by Vector Transform
Edoardo Mello Rella,Ajad Chhatkuli,Yun Liu,Ender Konukoglu,Luc Van Gool

Boundaries or contours are among the primary visual cues used by human and computer vision systems. One of the key problems in boundary detection is the loss formulation, which typically leads to class imbalance and, as a consequence, to thick boundaries which require non-differential post-processing steps to be thinned.

In this paper, we re-interpret boundaries as 1-D surfaces and formulate a one-to-one vector transform function that allows for training of boundary prediction completely avoiding the class imbalance issue. Specifically, we define the boundary representation at any point as the unit vector pointing to the closest boundary surface.

Our problem formulation leads to the estimation of direction as well as richer contextual information of the boundary, and, if desired, the availability of zero-pixel thin boundaries also at training time. Our method uses no hyper-parameter in the training loss and a fixed stable hyper-parameter at inference. We provide theoretical justification/discussions of the vector transform representation. We evaluate the proposed loss method using a standard architecture and show the excellent performance over other losses and representations on several datasets.
****

Efficient Training and Inference of Hypergraph Reasoning Networks
Guangxuan Xiao,Leslie Pack Kaelbling,Jiajun Wu,Jiayuan Mao

We study the problem of hypergraph reasoning in large domains, e.g., predicting the relationship between several entities based on the input facts. We observe that in logical reasoning, logical rules (e.g., my parent's parent is my grandparent) usually apply locally (e.g., only three people are involved in a grandparent rule), and sparsely (e.g., the grandparent relationship is sparse across all pairs of people in the world). Inspired by these observations, we propose Sparse and Local Neural Logic Machines (SpaLoc), a structured neural network for hypergraph reasoning. To leverage the sparsity in hypergraph neural networks, SpaLoc represents the grounding of relationships such as parent and grandparent as sparse tensors and uses neural networks and finite-domain quantification operations to infer new facts based on the input. We further introduce a sparsification loss to regularize the number of hyperedges in intermediate layers of a SpaLoc model. To enable training on large-scale graphs such as real-world knowledge graphs, SpaLoc makes training and inference-time sub-sampling of the input graphs. To remedy the information loss in sampled sub-graphs, we propose a novel sampling and label calibration paradigm based on an information-theoretic measure information sufficiency. Our SpaLoc shows superior accuracy and efficiency on synthetic datasets compared with prior art and achieves state-of-the-art performance on several real-world knowledge graph reasoning benchmarks.
****

Pixab-CAM: Attend Pixel, not Channel
Jaeeun Jang,Seokjun Kim,Hyeoncheol Kim

To understand the internal behaviors of convolution neural networks (CNNs), many class activation mapping (CAM) based methods, which generate an explanation map by a linear combination of channels and corresponding weights, have been proposed. Previous CAM-based methods have tried to define a channel-wise weight that represents the importance of a channel for the target class. However, these methods have two common limitations. First, all pixels in the channel share a single scalar value. If the pixels are tied to a specific value, some of them are overestimated. Second, since the explanation map is the result of a linear combination of channels in the activation tensor, it is inevitably dependent on the activation tensor. To address these issues, we propose gradient-free Pixel-wise Ablation-CAM (Pixab-CAM), which utilizes pixel-wise weights rather than channel-wise weights to break the link between pixels in a channel. In addition, in order not to generate an explanation map dependent on the activation tensor, the explanati on map is generated only with pixel-wise weights without linear combination with the activation tensor. In this paper, we also propose novel evaluation metrics to measure the quality of explanation maps using an adversarial attack. We demonstrate through experiments the qualitative and quantitative superiority of Pixab-CAM.

****

Distance-Based Background Class Regularization for Open-Set Recognition

Wonwoo Cho,Jaegul Choo

In open-set recognition (OSR), classifiers should be able to reject unknown-class samples while maintaining robust closed-set classification performance. To solve the OSR problem based on pre-trained Softmax classifiers, previous studies investigated offline analyses, e.g., distance-based sample rejection, which can limit the feature space of known-class data items. Since such classifiers are trained solely based on known-class samples, one can use background class regularization (BCR), which employs background-class data as surrogates of unknown-class data during training phase, to enhance OSR performance. However, previous regularization methods have limited OSR performance, since they categorized known-class data into a single group and then aimed to distinguish them from anomalies. In this paper, we propose a novel distance-based BCR method suitable for OSR, which limits the feature space of known-class data in a class-wise manner and then makes background-class samples located far away from the limited feature space. Instead of conventional Softmax classifiers, we use distance-based classifiers, which utilize the principle of linear discriminant analysis. Based on the distance measure used for classification, we design a novel regularization loss function that can contrast known-class and background-class samples while maintaining robust closed-set classification performance. Through our extensive experiments, we show that the proposed method provides robust OSR results with a simple inference process.

****

A Conditional Point Diffusion-Refinement Paradigm for 3D Point Cloud Completion

Zhaoyang Lyu,Zhifeng Kong,Xudong XU,Liang Pan,Dahua Lin

3D point clouds are an important data format that captures 3D information for real world objects. Since 3D point clouds scanned in the real world are often incomplete, it is important to recover the complete point cloud for many downstreaming applications. Most existing point cloud completion methods use the Chamfer Distance (CD) loss for training. The CD loss estimates correspondences between two point clouds by searching nearest neighbors, which does not capture the overall point distribution on the generated shape, and therefore likely leads to non-uniform point cloud generation. To tackle this problem, we propose a novel Point Diffusion-Refinement (PDR) paradigm for point cloud completion. PDR consists of a Conditional Generation Network (CGNet) and a ReFinement Network (RFNet). The CGNet uses a conditional generative model called the denoising diffusion probabilistic model (DDPM) to generate a coarse completion conditioned on the partial observation. DDPM establishes a one-to-one pointwise mapping between the generated point cloud and the uniform ground truth, and then optimizes the mean squared error loss to realize uniform generation. The RFNet refines the coarse output of the CGNet and further improves quality of the completed point cloud. In terms of the architecture, we develop a novel dual-path architecture for both networks. The architecture can (1) effectively and efficiently extract multi-level features from partially observed point clouds to guide completion, and (2) accurately manipulate spatial locations of 3D points to obtain smooth surfaces and sharp details. Extensive experimental results on various benchmark datasets show that our PDR paradigm outperforms previous state-of-the-art methods for point cloud completion. In addition, with the help of the RFNet, we can accelerate the iterative generation process of the DDPM by up to 50 times without much performance drop.

****

Auto-Transfer: Learning to Route Transferable Representations

Keerthiram Murugesan,Vijay Sadashivaiah,Ronny Luss,Karthikeyan Shanmugam,Pin-Yu Chen,Amit Dhurandhar

Knowledge transfer between heterogeneous source and target networks and tasks has received a lot of attention in recent times as large amounts of quality labeled data can be difficult to obtain in many applications. Existing approaches typically constrain the target deep neural network (DNN) feature representations to be close to the source DNNs feature representations, which can be limiting. We, in this paper, propose a novel adversarial multi-armed bandit approach that automatically learns to route source representations to appropriate target representations following which they are combined in meaningful ways to produce accurate target models. We see upwards of 5\% accuracy improvements compared with the state-of-the-art knowledge transfer methods on four benchmark (target) image datasets CUB200, Stanford Dogs, MIT67, and Stanford40 where the source dataset is ImageNet. We qualitatively analyze the goodness of our transfer scheme by showing individual examples of the important features focused on by our target network at different layers compared with the (closest) competitors. We also observe that our improvement over other methods is higher for smaller target datasets making it an effective tool for small data applications that may benefit from transfer learning.

****

Sharper Utility Bounds for Differentially Private Models

Yilin Kang,Yong Liu,Jian Li,Weipinng Wang

In this paper, by introducing Generalized Bernstein condition, we propose the first $\mathcal{O}\big(\frac{\sqrt{p}}{n\epsilon}\big)$ high probability excess population risk bound for differentially private algorithms under the assumptions G-Lipschitz, L-smooth, and Polyak-{\L}ojasiewicz condition, based on gradient perturbation method. If we replace the properties G-Lipschitz and L-smooth by α-H{\"o}lder smoothness (which can be used in non-smooth setting), the high probability bound comes to $\mathcal{O}\big(n^{-\frac{2\alpha}{1+2\alpha}}\big)$ w.r.t n, which cannot achieve $\mathcal{O}\left(1/n\right)$ when $\alpha\in(0,1]$. %and only better than previous results when $\alpha\in[1/2,1]$. To solve this problem, we propose a variant of gradient perturbation method, \textbf{max$\{1,g\}$-Normalized Gradient Perturbation} (m-NGP). We further show that by normalization, the high probability excess population risk bound under assumptions α-H{\"o}lder smooth and Polyak-{\L}ojasiewicz condition can achieve $\mathcal{O}\big(\frac{\sqrt{p}}{n\epsilon}\big)$, which is the first $\mathcal{O}\left(1/n\right)$ high probability utility bound w.r.t n for differentially private algorithms under non-smooth conditions. Moreover, we evaluate the performance of the new proposed algorithm m-NGP, the experimental results show that m-NGP improves the performance (measured by accuracy) of the DP model over real datasets. It demonstrates that m-NGP improves the excess population risk bound and the accuracy of the DP model on real datasets simultaneously.

****

Low-rank Matrix Recovery with Unknown Correspondence

Zhiwei Tang,Tsung-Hui Chang,Xiaojing Ye,Hongyuan Zha

We study a matrix recovery problem with unknown correspondence: given the observation matrix $M_o=[A,\tilde P B]$, where $\tilde P$ is an unknown permutation matrix, we aim to recover the underlying matrix $M=[A,B]$. Such problem commonly arises in many applications where heterogeneous data are utilized and the correspondence among them are unknown, e.g., due to privacy concerns. We show that it is possible to recover M via solving a nuclear norm minimization problem under a proper low-rank condition on M, with provable non-asymptotic error bound for the recovery of M. We propose an algorithm, M^3O (Matrix recovery via Min-Max Optimization) which recasts this combinatorial problem as a continuous minimax optimization problem and solves it by proximal gradient with a Max-Oracle. M^3O can also be applied to a more general scenario where we have missing entries in M_o and multiple groups of data with distinct unknown correspondence. Experiments on simulated data, the MovieLens 100K dataset and Yale B database show that M^3O achieves state-of-the-art performance over several baselines and can recover the ground-truth correspondence with high accuracy.

****

Multi-scale fusion self attention mechanism

Qibin Li,Nianmin Yao,Jian Zhao,Yanan Zhang

Self attention is widely used in various tasks because it can directly calculate the dependency between words, regardless of distance. However, the existing self attention lacks the ability to extract phrase level information. This is because the self attention only considers the one-to-one relationship between words and ignores the one-to-many relationship between words and phrases. Consequently, we design a multi-scale fusion self attention model for phrase information to resolve the above issues. Based on the traditional attention mechanism, multi-scale fusion self attention extracts phrase information at different scales by setting convolution kernels at different levels, and calculates the corresponding attention matrix at different scales, so that the model can better extract phrase level information. Compared with the traditional self attention model, we also designed a unique attention matrix sparsity strategy to better select the information that the model needs to pay attention to, so that our model can be more effective. Experimental results show that our model is superior to the existing baseline model in relation extraction task and GLUE task.

****

PoNet: Pooling Network for Efficient Token Mixing in Long Sequences

Chao-Hong Tan,Qian Chen,Wen Wang,Qinglin Zhang,Siqi Zheng,Zhen-Hua Ling

Transformer-based models have achieved great success in various NLP, vision, and speech tasks. However, the core of Transformer, the self-attention mechanism, has a quadratic time and memory complexity with respect to the sequence length, which hinders applications of Transformer-based models to long sequences. Many approaches have been proposed to mitigate this problem, such as sparse attention mechanisms, low-rank matrix approximations and scalable kernels, and token mixing alternatives to self-attention. We propose a novel Pooling Network (PoNet) for token mixing in long sequences with linear complexity. We design multi-granularity pooling and pooling fusion to capture different levels of contextual information and combine their interactions with tokens. On the Long Range Arena benchmark, PoNet significantly outperforms Transformer and achieves competitive accuracy, while being only slightly slower than the fastest model, FNet, across all sequence lengths measured on GPUs. We also conduct systematic studies on the transfer learning capability of PoNet and observe that PoNet achieves 95.7 percent of the accuracy of BERT on the GLUE benchmark, outperforming FNet by 4.5 percent relative. Comprehensive ablation analysis demonstrates effectiveness of the designed multi-granularity pooling and pooling fusion for token mixing in long sequences and efficacy of the designed pre-training tasks for PoNet to learn transferable contextualized language representations.

****

Connecting Data to Mechanisms with Meta Structual Causal Model

Gong Heyang

Recent years have seen impressive progress in theoretical and algorithmic developments of causal inference across various disciplines in science and engineering. However, there is still some unresolved theoretical problems, especially for cyclic causal relationships. In this article, we propose a meta structure causal model (Meta-SCM) framework inspired by understanding causality as information transfer. A key feature of our framework is the introduction of the concept of \emph{active mechanisms} to connect data and the collection of underlying causal mechanisms. We show that the Meta-SCM provides a novel approach to address the theoretical complications for modeling cyclic causal relations. In addition, we propose a \emph{sufficient activated mechanisms} assumption, and explain its relationship with existing assumptions in causal representation learning. Finally, we conclude the main idea of the meta-SCM framework with an emphasis on its theoretical and conceptual novelty.

****

Huber Additive Models for Non-stationary Time Series Analysis

Yingjie Wang,Xianrui Zhong,Fengxiang He,Hong Chen,Dacheng Tao

Sparse additive models have shown promising flexibility and interpretability in processing time series data. However, existing methods usually assume the time se ries data to be stationary and the innovation is sampled from a Gaussian distribution. Both assumptions are too stringent for heavy-tailed and non-stationary time series data that frequently arise in practice, such as finance and medical fields. To address these problems, we propose an adaptive sparse Huber additive model for robust forecasting in both non-Gaussian data and (non)stationary data. In theory, the generalization bounds of our estimator are established for both stationary and nonstationary time series data, which are independent of the widely used mixing conditions in learning theory of dependent observations. Moreover, the error bound for non-stationary time series contains a discrepancy measure for the shifts of the data distributions over time. Such a discrepancy measure can be estimated empirically and used as a penalty in our method. Experimental results on both synthetic and real-world benchmark datasets validate the effectiveness of the proposed method. The code is available at https://github.com/xianruizhong/SpHAM.

****

Heterologous Normalization

Chunjie Luo,Jianfeng Zhan,Lei Wang,Wanling Gao

Batch Normalization has become a standard technique for training modern deep networks. However, its effectiveness diminishes when the batch size becomes smaller since the batch statistics estimation becomes inaccurate. This paper proposes Heterologous Normalization, which computes normalization's mean and standard deviation from different pixel sets to take advantage of different normalization methods. Specifically, it calculates the mean like Batch Normalization to maintain the advantage of Batch Normalization. Meanwhile, it enlarges the number of pixels from which the standard deviation is calculated, thus alleviating the problem caused by the small batch size. Experiments show that Heterologous Normalization surpasses or achieves comparable performance to existing homologous methods, with large or small batch sizes on various datasets.

****

Model-augmented Prioritized Experience Replay

Youngmin Oh,Jinwoo Shin,Eunho Yang,Sung Ju Hwang

Experience replay is an essential component in off-policy model-free reinforcement learning (MfRL). Due to its effectiveness, various methods for calculating priority scores on experiences have been proposed for sampling. Since critic networks are crucial to policy learning, TD-error, directly correlated to Q-values, is one of the most frequently used features to compute the scores. However, critic networks often under- or overestimate Q-values, so it is often ineffective to learn to predict Q-values by sampled experiences based heavily on TD-error. Accordingly, it is valuable to find auxiliary features, which positively support TD-error in calculating the scores for efficient sampling. Motivated by this, we propose a novel experience replay method, which we call model-augmented prioritized experience replay (MaPER), that employs new learnable features driven from components in model-based RL (MbRL) to calculate the scores on experiences. The proposed MaPER brings the effect of curriculum learning for predicting Q-values better by the critic network with negligible memory and computational overhead compared to the vanilla PER. Indeed, our experimental results on various tasks demonstrate that MaPER can significantly improve the performance of the state-of-the-art off-policy MfRL and MbRL which includes off-policy MfRL algorithms in its policy optimization procedure.

****

GANet: Glyph-Attention Network for Few-Shot Font Generation

Mingtao Guo,Wei Xiong,Zheng Wang,Yong Tang,Ting Wu

Font generation is a valuable but challenging task, it is time consuming and costly to design font libraries which cover all glyphs with various styles. The time and cost of such task will be greatly reduced if the complete font library can be generated from only a few custom samples. Inspired by font characteristics and global and local attention mechanism Wang et al. (2018), we propose a glyph-attention network (GANet) to tackle this problem. Firstly, a content encoder and a style encoder are trained to extract features as keys and values from a content glyph set and a style glyph set, respectively. Secondly, a query vector genera ted from a single glyph sample by the query encoder is applied to draw out proper features from the content and style (key, value) pairs via glyph-attention modules. Next, a decoder is used to recover a glyph from the queried features. Lastly, Adversarial losses Goodfellow et al. (2014) with multi-task glyph discriminator are employed to stablize the training process. Experimental results demonstrate that our method is able to create robust results with superior fidelity. Less number of samples are needed and better performance is achieved when compared to the other state-of-the-art few-shot font generation methods, without utilizing supervision on locality such as component, skeleton, or strokes, etc.

****

Post-Training Detection of Backdoor Attacks for Two-Class and Multi-Attack Scenarios

Zhen Xiang,David Miller,George Kesidis

Backdoor attacks (BAs) are an emerging threat to deep neural network classifiers. A victim classifier will predict to an attacker-desired target class whenever a test sample is embedded with the same backdoor pattern (BP) that was used to poison the classifier's training set. Detecting whether a classifier is backdoor attacked is not easy in practice, especially when the defender is, e.g., a downstream user without access to the classifier's training set. This challenge is addressed here by a reverse-engineering defense (RED), which has been shown to yield state-of-the-art performance in several domains. However, existing REDs are not applicable when there are only two classes or when multiple attacks are present. These scenarios are first studied in the current paper, under the practical constraints that the defender neither has access to the classifier's training set nor to supervision from clean reference classifiers trained for the same domain. We propose a detection framework based on BP reverse-engineering and a novel expected transferability (ET) statistic. We show that our ET statistic is effective using the same detection threshold, irrespective of the classification domain, the attack configuration, and the BP reverse-engineering algorithm that is used. The excellent performance of our method is demonstrated on six benchmark datasets. Notably, our detection framework is also applicable to multi-class scenarios with multiple attacks. Code is available at https://github.com/zhenxianglance/2ClassBADetection.

****

Multi-Task Processes

Donggyun Kim,Seongwoong Cho,Wonkwang Lee,Seunghoon Hong

Neural Processes (NPs) consider a task as a function realized from a stochastic process and flexibly adapt to unseen tasks through inference on functions. However, naive NPs can model data from only a single stochastic process and are designed to infer each task independently. Since many real-world data represent a set of correlated tasks from multiple sources (e.g., multiple attributes and multi-sensor data), it is beneficial to infer them jointly and exploit the underlying correlation to improve the predictive performance.

To this end, we propose Multi-Task Neural Processes (MTNPs), an extension of NPs designed to jointly infer tasks realized from multiple stochastic processes. We build MTNPs in a hierarchical way such that inter-task correlation is considered by conditioning all per-task latent variables on a single global latent variable. In addition, we further design our MTNPs so that they can address multi-task settings with incomplete data (i.e., not all tasks share the same set of input points), which has high practical demands in various applications.

Experiments demonstrate that MTNPs can successfully model multiple tasks jointly by discovering and exploiting their correlations in various real-world data such as time series of weather attributes and pixel-aligned visual modalities. We release our code at https://github.com/GitGyun/multi_task_neural_processes.

****

Reward Shifting for Optimistic Exploration and Conservative Exploitation

Hao Sun,Lei Han,Jian Guo,Bolei Zhou

In this work, we study the simple yet universally applicable case of reward shaping, the linear transformation, in value-based Deep Reinforcement Learning. We show that reward shifting, as the simplest linear reward transformation, is equiv

alent to changing 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. In this case, a conservative exploitation improves offline RL value estimation, and the optimistic value estimation benefits the exploration of online RL. We verify our insight on a range of tasks: (1) In offline RL, the conservative exploitation leads to improved learning performance based on off-the-shelf algorithms; (2) In online continuous control, multiple value functions with different shifting constants can be used to trade-off between exploration and exploitation thus improving learning efficiency; (3) In online RL with discrete action space, a negative reward shifting brings an improvement over the previous curiosity-based exploration method.
****
Information-Aware Time Series Meta-Contrastive Learning
Dongsheng Luo,Wei Cheng,Yingheng Wang,Dongkuan Xu,Jingchao Ni,Wenchao Yu,Xuchao Zhang,Yanchi Liu,Haifeng Chen,Xiang Zhang
Various contrastive learning approaches have been proposed in recent years and achieve significant empirical success. While effective and prevalent, contrastive learning has been less explored for time series data. A key component of contrastive learning is to select appropriate augmentations imposing some priors to construct feasible positive samples, such that an encoder can be trained to learn robust and discriminative representations. Unlike image and language domains where ``desired'' augmented samples can be generated with the rule of thumb guided by prefabricated human priors, the ad-hoc manual selection of time series augmentations is hindered by their diverse and human-unrecognizable temporal structures. How to find the desired augmentations of time series data that are meaningful for given contrastive learning tasks and datasets remains an open question. In this work, we address the problem by encouraging both high fidelity and variety based upon information theory. A theoretical analysis leads to the criteria for selecting feasible data augmentations. On top of that, we employ the meta-learning mechanism and propose an information-aware approach, InfoTS, that adaptively selects optimal time series augmentations for contrastive representation learning. The meta-learner and the encoder are jointly optimized in an end-to-end manner to avoid sub-optimal solutions. Experiments on various datasets show highly competitive performance with up to 11.4% reduction in MSE on the forecasting task and up to 2.8% relative improvement in accuracy on the classification task over the leading baselines.
****
Dynamic Token Normalization improves Vision Transformers
Wenqi Shao,Yixiao Ge,Zhaoyang Zhang,XUYUAN XU,Xiaogang Wang,Ying Shan,Ping Luo
Vision Transformer (ViT) and its variants (e.g., Swin, PVT) have achieved great success in various computer vision tasks, owing to their capability to learn long-range contextual information. Layer Normalization (LN) is an essential ingredient in these models. However, we found that the ordinary LN makes tokens at different positions similar in magnitude because it normalizes embeddings within each token. It is difficult for Transformers to capture inductive bias such as the positional context in an image with LN. We tackle this problem by proposing a new normalizer, termed Dynamic Token Normalization (DTN), where normalization is performed both within each token (intra-token) and across different tokens (inter-token). DTN has several merits. Firstly, it is built on a unified formulation and thus can represent various existing normalization methods. Secondly, DTN learns to normalize tokens in both intra-token and inter-token manners, enabling Transformers to capture both the global contextual information and the local positional context. Thirdly, by simply replacing LN layers, DTN can be readily plugged into various vision transformers, such as ViT, Swin, and PVT. Extensive experiments show that the transformer equipped with DTN consistently outperforms baseline model with minimal extra parameters and computational overhead. For example, DTN outperforms LN on small ViT by 1.1% top-1 accuracy on ImageNet.
****
Symbolic Learning to Optimize: Towards Interpretability and Scalability

Wenqing Zheng,Tianlong Chen,Ting-Kuei Hu,Zhangyang Wang

Recent studies on Learning to Optimize (L2O) suggest a promising path to automating and accelerating the optimization procedure for complicated tasks. Existing L2O models parameterize optimization rules by neural networks, and learn those numerical rules via meta-training. However, they face two common pitfalls: (1) scalability: the numerical rules represented by neural networks create extra memory overhead for applying L2O models, and limits their applicability to optimizing larger tasks; (2) interpretability: it is unclear what each L2O model has learned in its black-box optimization rule, nor is it straightforward to compare different L2O models in an explainable way. To avoid both pitfalls, this paper proves the concept that we can "kill two birds by one stone", by introducing the powerful tool of symbolic regression to L2O. In this paper, we establish a holistic symbolic representation and analysis framework for L2O, which yields a series of insights for learnable optimizers. Leveraging our findings, we further propose a lightweight L2O model that can be meta-trained on large-scale problems and outperformed human-designed and tuned optimizers. Our work is set to supply a brand-new perspective to L2O research. Codes are available at: https://github.com/VITA-Group/Symbolic-Learning-To-Optimize.

****

Reasoning With Hierarchical Symbols: Reclaiming Symbolic Policies For Visual Reinforcement Learning

Wenqing Zheng,S P Sharan,Zhiwen Fan,Zhangyang Wang

Deep vision models are nowadays widely integrated into visual reinforcement learning (RL) to parameterize the policy networks. However, the learned policies are overparameterized black boxes that lack interpretability, and are usually brittle under input distribution shifts. This work revisits this end-to-end learning pipeline, and proposes an alternative stage-wise approach that features hierarchical reasoning. Specifically, our approach progressively converts a policy network into the interpretable symbolic policy, composed from geometric and numerical symbols and operators. A policy regression algorithm called RoundTourMix is proposed to distill the symbolic rules as teacher-student. The symbolic policy can be treated as discrete and abstracted representations of the policy network, but are found to be more interpretable, robust and transferable. The proposed symbolic distillation approach is experimentally demonstrated to maintain the performance and ``de-noise" the CNN policy: on six specific environments, our distilled symbolic policy achieved compelling or even higher scores than the CNN based RL agents. Our codes will be fully released upon acceptance.

****

PhaseFool: Phase-oriented Audio Adversarial Examples via Energy Dissipation

Ziyue Jiang,Yi Ren,Zhou Zhao

Audio adversarial attacks design perturbations onto inputs that lead an automatic speech recognition (ASR) model to predict incorrect outputs. Current audio adversarial attacks optimize perturbations with different constraints (e.g. lp-norm for waveform or the principle of auditory masking for magnitude spectrogram) to achieve their imperceptibility. Since phase is not relevant for speech recognition, the existing audio adversarial attacks neglect the influence of phase spectrogram. In this work, we propose a novel phase-oriented algorithm named PhaseFool that can efficiently construct imperceptible audio adversarial examples with energy dissipation. Specifically, we leverage the spectrogram consistency of short-time Fourier transform (STFT) to adversarially transfer phase perturbations to the adjacent frames of magnitude spectrogram and dissipate the energy that is crucial for ASR systems. Moreover, we propose a weighted loss function to improve the imperceptibility of PhaseFool. Experimental results demonstrate that PhaseFool can inherently generate full-sentence imperceptible audio adversarial examples with the 100% targeted success rate within 500 steps on average (9.24x speed-up over current state-of-the-art imperceptible counterparts), which is verified through a human study. Most importantly, our PhaseFool is the first to exploit the phase-oriented energy dissipation in the audio adversarial examples rather than add perturbations on the audio waveform like most previous works.

****

Resolving Training Biases via Influence-based Data Relabeling
Shuming Kong,Yanyan Shen,Linpeng Huang

The performance of supervised learning methods easily suffers from the training bias issue caused by train-test distribution mismatch or label noise. Influence function is a technique that estimates the impacts of a training sample on the model’s predictions. Recent studies on \emph{data resampling} have employed influence functions to identify \emph{harmful} training samples that will degrade model's test performance. They have shown that discarding or downweighting the identified harmful training samples is an effective way to resolve training biases. In this work, we move one step forward and propose an influence-based relabeling framework named RDIA for reusing harmful training samples toward better model performance. To achieve this, we use influence functions to estimate how relabeling a training sample would affect model's test performance and further develop a novel relabeling function R. We theoretically prove that applying R to relabel harmful training samples allows the model to achieve lower test loss than simply discarding them for any classification tasks using cross-entropy loss. Extensive experiments on ten real-world datasets demonstrate RDIA outperforms the state-of-the-art data resampling methods and improves model's robustness against label noise.

****

Modeling Adversarial Noise for Adversarial Defense
Dawei Zhou,Nannan Wang,Bo Han,Tongliang Liu

Deep neural networks have been demonstrated to be vulnerable to adversarial noise, promoting the development of defense against adversarial attacks. Motivated by the fact that adversarial noise contains well-generalizing features and that the relationship between adversarial data and natural data can help infer natural data and make reliable predictions, in this paper, we study to model adversarial noise by learning the transition relationship between adversarial labels (i.e. the flipped labels used to generate adversarial data) and natural labels (i.e. the ground truth labels of the natural data). Specifically, we introduce an instance-dependent transition matrix to relate adversarial labels and natural labels, which can be seamlessly embedded with the target model (enabling us to model stronger adaptive adversarial noise). Empirical evaluations demonstrate that our method could effectively improve adversarial accuracy.

****

Sequential Reptile: Inter-Task Gradient Alignment for Multilingual Learning
Seanie Lee,Hae Beom Lee,Juho Lee,Sung Ju Hwang

Multilingual models jointly pretrained on multiple languages have achieved remarkable performance on various multilingual downstream tasks. Moreover, models finetuned on a single monolingual downstream task have shown to generalize to unseen languages. In this paper, we first show that it is crucial for those tasks to align gradients between them in order to maximize knowledge transfer while minimizing negative transfer. Despite its importance, the existing methods for gradient alignment either have a completely different purpose, ignore inter-task alignment, or aim to solve continual learning problems in rather inefficient ways. As a result of the misaligned gradients between tasks, the model suffers from severe negative transfer in the form of catastrophic forgetting of the knowledge acquired from the pretraining. To overcome the limitations, we propose a simple yet effective method that can efficiently align gradients between tasks. Specifically, we perform each inner-optimization by sequentially sampling batches from all the tasks, followed by a Reptile outer update. Thanks to the gradients aligned between tasks by our method, the model becomes less vulnerable to negative transfer and catastrophic forgetting. We extensively validate our method on various multi-task learning and zero-shot cross-lingual transfer tasks, where our method largely outperforms all the relevant baselines we consider.

****

Representational Continuity for Unsupervised Continual Learning
Divyam Madaan,Jaehong Yoon,Yuanchun Li,Yunxin Liu,Sung Ju Hwang

Continual learning (CL) aims to learn a sequence of tasks without forgetting the previously acquired knowledge. However, recent CL advances are restricted to supervised continual learning (SCL) scenarios. Consequently, they are not scalable to real-world applications where the data distribution is often biased and unannotated. In this work, we focus on unsupervised continual learning (UCL), where we learn the feature representations on an unlabelled sequence of tasks and show that reliance on annotated data is not necessary for continual learning. We conduct a systematic study analyzing the learned feature representations and show that unsupervised visual representations are surprisingly more robust to catastrophic forgetting, consistently achieve better performance, and generalize better to out-of-distribution tasks than SCL. Furthermore, we find that UCL achieves a smoother loss landscape through qualitative analysis of the learned representations and learns meaningful feature representations. Additionally, we propose Lifelong Unsupervised Mixup (LUMP), a simple yet effective technique that interpolates between the current task and previous tasks' instances to alleviate catastrophic forgetting for unsupervised representations.

****

Pseudo Numerical Methods for Diffusion Models on Manifolds

Luping Liu,Yi Ren,Zhijie Lin,Zhou Zhao

Denoising Diffusion Probabilistic Models (DDPMs) can generate high-quality samples such as image and audio samples. However, DDPMs require hundreds to thousands of iterations to produce a sample. Several prior works have successfully accelerated DDPMs through adjusting the variance schedule (e.g., Improved Denoising Diffusion Probabilistic Models) or the denoising equation (e.g., Denoising Diffusion Implicit Models (DDIMs)). However, these acceleration methods cannot maintain the quality of samples and even introduce new noise at high speedup rate, which limit their practicability. To accelerate the inference process while keeping the sample quality, we provide a new perspective that DDPMs should be treated as solving differential equations on manifolds. Under such a perspective, we propose pseudo numerical methods for diffusion models (PNDMs). Specifically, we figure out how to solve differential equations on manifolds and show that DDIMs are simple cases of pseudo numerical methods. We change several classical numerical methods to corresponding pseudo numerical methods and find that pseudo linear multi-step method is the best method in most situations. According to our experiments, by directly using pre-trained models on Cifar10, CelebA and LSUN, PNDMs can generate higher quality synthetic images with only 50 steps compared with 1000-step DDIMs (20x speedup), significantly outperform DDIMs with 250 steps (by around 0.4 in FID) and have good generalization on different variance schedules.

****

Reinforcement Learning for Adaptive Mesh Refinement

Jiachen Yang,Tarik Dzanic,Brenden K. Petersen,Jun Kudo,Ketan Mittal,Jean-Sylvain Camier,Vladimir Tomov,Tuo Zhao,Hongyuan Zha,Tzanio Kolev,Robert Anderson,Daniel faissol

Large-scale finite element simulations of complex physical systems governed by partial differential equations (PDE) crucially depend on adaptive mesh refinement (AMR) to allocate computational budget to regions where higher resolution is required. Existing scalable AMR methods make heuristic refinement decisions based on instantaneous error estimation and thus do not aim for long-term optimality over an entire simulation. We propose a novel formulation of AMR as a Markov decision process and apply deep reinforcement learning (RL) to train refinement {\it policies} directly from simulation. AMR poses a new problem for RL as both the state dimension and available action set changes at every step, which we solve by proposing new policy architectures with differing generality and inductive bias. The model sizes of these policy architectures are independent of the mesh size and hence can be deployed on larger simulations than those used at train time. We demonstrate in comprehensive experiments on static function estimation and time-dependent equations that RL policies can be trained on problems without using ground truth solutions, are competitive with a widely-used error estimator, and generalize to larger, more complex, and unseen test problems.

****

Supervision Exists Everywhere: A Data Efficient Contrastive Language-Image Pre-training Paradigm

Yangguang Li,Feng Liang,Lichen Zhao,Yufeng Cui,Wanli Ouyang,Jing Shao,Fengwei Yu,Junjie Yan

Recently, large-scale Contrastive Language-Image Pre-training (CLIP) has attracted unprecedented attention for its impressive zero-shot recognition ability and excellent transferability to downstream tasks. However, CLIP is quite data-hungry and requires 400M image-text pairs for pre-training, thereby restricting its adoption. This work proposes a novel training paradigm, Data efficient CLIP (DeCLIP), to alleviate this limitation. We demonstrate that by carefully utilizing the widespread supervision among the image-text pairs, our De-CLIP can learn generic visual features more efficiently. Instead of using the single image-text contrastive supervision, we fully exploit data potential through the use of (1) self-supervision within each modality; (2) multi-view supervision across modalities; (3) nearest-neighbor supervision from other similar pairs. Benefiting from intrinsic supervision, our DeCLIP-ResNet50 can achieve 60.4% zero-shot top1 accuracy on ImageNet, which is 0.8% above the CLIP-ResNet50 while using 7.1×fewer data. Our DeCLIP-ResNet50 outperforms its counterpart in 8 out of 11 visual datasets when transferred to downstream tasks. Moreover, Scaling up the model and computing also works well in our framework.

****

Environment Predictive Coding for Visual Navigation

Santhosh Kumar Ramakrishnan,Tushar Nagarajan,Ziad Al-Halah,Kristen Grauman

We introduce environment predictive coding, a self-supervised approach to learn environment-level representations for embodied agents. In contrast to prior work on self-supervised learning for individual images, we aim to encode a 3D environment using a series of images observed by an agent moving in it. We learn these representations via a masked-zone prediction task, which segments an agent’s trajectory into zones and then predicts features of randomly masked zones, conditioned on the agent’s camera poses. This explicit spatial conditioning encourages learning representations that capture the geometric and semantic regularities of 3D environments. We learn such representations on a collection of video walkthroughs and demonstrate successful transfer to multiple downstream navigation tasks. Our experiments on the real-world scanned 3D environments of Gibson and Matterport3D show that our method obtains 2 - 6× higher sample-efficiency and up to 57% higher performance over standard image-representation learning.

****

Momentum Doesn't Change The Implicit Bias

Bohan Wang,Qi Meng,Huishuai Zhang,Ruoyu Sun,Wei Chen,Zhi-Ming Ma

The momentum acceleration technique is widely adopted in many optimization algorithms. However, the theoretical understanding on how the momentum affects the generalization performance of the optimization algorithms is still unknown. In this paper, we answer this question through analyzing the implicit bias of momentum-based optimization. We prove that both SGD with momentum and Adam converge to the L_2 max-margin solution for exponential-tailed loss, which is the same as vanilla gradient descent.

That means, these optimizers with momentum acceleration still converge to a model with low complexity, which provides guarantees on their generalization. Technically, to overcome the difficulty brought by the error accumulation in analyzing the momentum, we construct new Lyapunov functions as a tool to analyze the gap between the model parameter and the max-margin solution.

****

ZARTS: On Zero-order Optimization for Neural Architecture Search

Xiaoxing Wang,Wenxuan Guo,Junchi Yan,Xiaokang Yang,Jianlin Su

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 will often distort the loss landscape, leading to the biased objective to optimize and in turn inaccurate gradient estimation for architec ture 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, which are state-of-the-art performance.

****

Topological Experience Replay

Zhang-Wei Hong,Tao Chen,Yen-Chen Lin,Joni Pajarinen,Pulkit Agrawal

State-of-the-art deep Q-learning methods update Q-values using state transition tuples sampled from the experience replay buffer. This strategy often randomly samples or prioritizes data sampling based on measures such as the temporal difference (TD) error. Such sampling strategies can be inefficient at learning Q-function since a state's correct Q-value preconditions on the accurate successor states' Q-value. Disregarding such a successor's value dependency leads to useless updates and even learning wrong values.

To expedite Q-learning, we maintain states' dependency by organizing the agent's experience into a graph. Each edge in the graph represents a transition between two connected states. We perform value backups via a breadth-first search that expands vertices in the graph starting from the set of terminal states successively moving backward. We empirically show that our method is substantially more data-efficient than several baselines on a diverse range of goal-reaching tasks. Notably, the proposed method also outperforms baselines that consume more batches of training experience.

****

Sparsity Winning Twice: Better Robust Generalization from More Efficient Training

Tianlong Chen,Zhenyu Zhang,pengjun wang,Santosh Balachandra,Haoyu Ma,Zehao Wang, Zhangyang Wang

Recent studies demonstrate the deep networks, even robustified by the state-of-the-art adversarial training (AT), still suffer from large robust generalization gaps, in addition to the much more expensive training costs than standard training. In this paper, we investigate this intriguing problem from a new perspective, i.e., $\textit{injecting appropriate forms of sparsity}$ during adversarial training. We introduce two alternatives for sparse adversarial training: (i) $\textit{static sparsity}$, by leveraging recent results from the lottery ticket hypothesis to identify critical sparse subnetworks arising from the early training; (ii) $\textit{dynamic sparsity}$, by allowing the sparse subnetwork to adaptively adjust its connectivity pattern (while sticking to the same sparsity ratio) throughout training. We find both static and dynamic sparse methods to yield win-win: substantially shrinking the robust generalization gap and alleviating the robust overfitting, meanwhile significantly saving training and inference FLOPs. Extensive experiments validate our proposals with multiple network architectures on diverse datasets, including CIFAR-10/100 and Tiny-ImageNet. For example, our methods reduce robust generalization gap and overfitting by 34.44% and 4.02%, with comparable robust/standard accuracy boosts and 87.83%/87.82% training/inference FLOPs savings on CIFAR-100 with ResNet-18. Besides, our approaches can be organically combined with existing regularizers, establishing new state-of-the-art results in AT. All codes are included.

****

CrossMatch: Cross-Classifier Consistency Regularization for Open-Set Single Domain Generalization

Ronghang Zhu,Sheng Li

Single domain generalization (SDG) is a challenging scenario of domain generalization, where only one source domain is available to train the model. Typical SDG methods are based on the adversarial data augmentation strategy, which complements the diversity of source domain to learn a robust model. Existing SDG methods require the source and target domains to have the same label space. However, as target domains may contain novel categories unseen in source label space, this assumption is not practical in many real-world applications. In this paper, we propose a challenging and untouched problem: \textit{Open-Set Single Domain Generalization} (OS-SDG), where target domains include unseen categories out of source label space. The goal of OS-SDG is to learn a model, with only one source domain, to classify a target sample with correct class if it belongs to source label space, or assign it to unknown classes. We design a \textit{CrossMatch} approach to improve the performance of SDG methods on identifying unknown classes by leveraging a multi-binary classifier. CrossMatch generates auxiliary samples out of source label space by using an adversarial data augmentation strategy. We also adopt a consistency regularization on generated auxiliary samples between multi-binary classifiers and the model trained by SDG methods, to improve the model’s capability on unknown class identification. Experimental results on benchmark datasets prove the effectiveness of CrossMatch on enhancing the performance of SDG methods in the OS-SDG setting.
****

A Variance Principle Explains why Dropout Finds Flatter Minima
Zhongwang Zhang,Hanxu Zhou,Zhiqin Xu
Although dropout has achieved great success in deep learning, little is known about how it helps the training find a good generalization solution in the high-dimensional parameter space. In this work, we show that the training with dropout finds the neural network with a flatter minimum compared with standard gradient descent training. We further study the underlying mechanism of why dropout finds flatter minima through experiments. We propose a Variance Principle that the variance of a noise is larger at the sharper direction of the loss landscape. Existing works show that SGD satisfies the variance principle, which leads the training to flatter minima. Our work show that the noise induced by the dropout also satisfies the variance principle that explains why dropout finds flatter minima. In general, our work points out that the variance principle is an important similarity between dropout and SGD that lead the training to find flatter minima and obtain good generalization.
****

Robust Unlearnable Examples: Protecting Data Privacy Against Adversarial Learning
Shaopeng Fu,Fengxiang He,Yang Liu,Li Shen,Dacheng Tao
The tremendous amount of accessible data in cyberspace face the risk of being unauthorized used for training deep learning models. To address this concern, methods are proposed to make data unlearnable for deep learning models by adding a type of error-minimizing noise. However, such conferred unlearnability is found fragile to adversarial training. In this paper, we design new methods to generate robust unlearnable examples that are protected from adversarial training. We first find that the vanilla error-minimizing noise, which suppresses the informative knowledge of data via minimizing the corresponding training loss, could not effectively minimize the adversarial training loss. This explains the vulnerability of error-minimizing noise in adversarial training. Based on the observation, robust error-minimizing noise is then introduced to reduce the adversarial training loss. Experiments show that the unlearnability brought by robust error-minimizing noise can effectively protect data from adversarial training in various scenarios. The code is available at \url{https://github.com/fshp971/robust-unlearnable-examples}.
****

Log-Polar Space Convolution
Bing Su,Ji-Rong Wen
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) method, 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.

****

A NON-PARAMETRIC REGRESSION VIEWPOINT : GENERALIZATION OF OVERPARAMETRIZED DEEP RELU NETWORK UNDER NOISY OBSERVATIONS

Namjoon Suh,Hyunouk Ko,Xiaoming Huo

We study the generalization properties of the overparameterized deep neural network (DNN) with Rectified Linear Unit (ReLU) activations.

Under the non-parametric regression framework, it is assumed that the ground-truth function is from a reproducing kernel Hilbert space (RKHS) induced by a neural tangent kernel (NTK) of ReLU DNN, and a dataset is given with the noises. Without a delicate adoption of early stopping, we prove that the overparametrized DNN trained by vanilla gradient descent does not recover the ground-truth function. It turns out that the estimated DNN's L_{2} prediction error is bounded away from 0. As a complement of the above result, we show that the ℓ_{2}-regularized gradient descent enables the overparametrized DNN achieve the minimax optimal convergence rate of the L_{2} prediction error, without early stopping. Notably, the rate we obtained is faster than $\mathcal{O}(n^{-1/2})$ known in the literature.

****

DSEE: Dually Sparsity-embedded Efficient Tuning of Pre-trained Language Models

Xuxi Chen,Tianlong Chen,Yu Cheng,Weizhu Chen,Zhangyang Wang,Ahmed Hassan Awadallah

Gigantic pre-trained models have become central to natural language processing (NLP), serving as the starting point for fine-tuning towards a range of downstream tasks. However, two pain points persist for this paradigm: (a) as the pre-trained models grow bigger (e.g., 175B parameters for GPT-3), even the fine-tuning process can be time-consuming and computationally expensive; (b) the fine-tuned model has the same size as its starting point by default, which is neither sensible due to its more specialized functionality, nor practical since many fine-tuned models will be deployed in resource-constrained environments. To address these pain points, we propose a framework for resource- and parameter-efficient fine-tuning by leveraging the sparsity prior in both weight updates and the final model weights. Our proposed framework, dubbed $\textbf{D}$ually $\textbf{S}$parsity-$\textbf{E}$mbedded $\textbf{E}$fficient Tuning (DSEE), aims to achieve two key objectives: (i) $\textit{parameter efficient fine-tuning}$ - by enforcing sparsity-aware weight updates on top of the pre-trained weights; and (ii) $\textit{resource-efficient inference}$ - by encouraging a sparse weight structure towards the final fine-tuned model. We leverage sparsity in these two directions by exploiting both unstructured and structural sparse patterns in pre-trained language models via magnitude-based pruning and ℓ_1 sparse regularization. Extensive experiments and in-depth investigations, with diverse network backbones (i.e., BERT, GPT-2, and DeBERTa) on dozens of datasets, consistently demonstrate highly impressive parameter-/training-/inference-efficiency, while maintaining competitive downstream transfer performance. For instance, our DSEE-BERT obtains about 35% inference FLOPs savings with $<0.1\%$ trainable parameters and comparable performance to conventional fine-tuning.

****

Disentangling Properties of Contrastive Methods

Jinkun Cao,Qing Yang,Jialei Huang,Yang Gao

Disentangled representation learning is an important topic in representation learning, since it not only allows the representation to be human interpretable, but it is also robust and benefits downstream task performance. Prior methods achieved initial successes on simplistic synthetic datasets but failed to scale to complex real-world datasets. Most of the previous methods adopt image generative models, such as GAN and VAE, to learn the disentangled representation. But we observe they are hard to learn disentangled representation on real-world images. Recently, self-supervised contrastive methods such as MoCo, SimCLR, and BYOL have achieved impressive performances on large-scale visual recognition tasks. In this paper, we explored the possibility of using contrastive methods to learn a disentangled representation, a discriminative approach that is drastically different from previous approaches. Surprisingly, we find that the contrastive method learns a disentangled representation with only minor modifications. The contrastively learned representation satisfies a ``group disentanglement'' property, which is a relaxed version of the original disentanglement property. This relaxation might be useful for scaling disentanglement learning to large and complex datasets. We further find contrastive methods achieve state-of-thet-art disentanglement performance on several widely used benchmarks, such as dSprites and Car3D. It also achieves significantly higher performance on the real-world dataset CelebA.

****

Decision boundary variability and generalization in neural networks
Shiye Lei,Fengxiang He,Yancheng Yuan,Dacheng Tao

Existing works suggest that the generalizability is guaranteed when the margin between data and decision boundaries is sufficiently large. However, the existence of adversarial examples in neural networks shows that excellent generalization and small margin can exist simultaneously, which casts shadows to the current understanding. This paper discovers that the neural network with lower decision boundary (DB) variability has better generalizability. Two new notions, algorithm DB variability and (ϵ, η)-data DB variability, are proposed to measure the decision boundary variability from the algorithm and data perspectives. Extensive experiments show significant negative correlations between the decision boundary variability and the generalizability. From the theoretical view, we prove two lower bounds and two upper bounds of the generalization error based on the decision boundary variability, which is consistent with our empirical results. Moreover, the bounds do not explicitly depend on the network size, which is usually prohibitively large in deep learning.

****

Programmatic Reinforcement Learning without Oracles
Wenjie Qiu,He Zhu

Deep reinforcement learning (RL) has led to encouraging successes in many challenging control tasks. However, a deep RL model lacks interpretability due to the difficulty of identifying how the model's control logic relates to its network structure. Programmatic policies structured in more interpretable representations emerge as a promising solution. Yet two shortcomings remain: First, synthesizing programmatic policies requires optimizing over the discrete and non-differentiable search space of program architectures. Previous works are suboptimal because they only enumerate program architectures greedily guided by a pretrained RL oracle. Second, these works do not exploit compositionality, an important programming concept, to reuse and compose primitive functions to form a complex function for new tasks. Our first contribution is a programmatically interpretable RL framework that conducts program architecture search on top of a continuous relaxation of the architecture space defined by programming language grammar rules. Our algorithm allows policy architectures to be learned with policy parameters via bilevel optimization using efficient policy-gradient methods, and thus does not require a pretrained oracle. Our second contribution is improving programmatic policies to support compositionality by integrating primitive functions learned to grasp task-agnostic skills as a composite program to solve novel RL problems. Experiment results demonstrate that our algorithm excels in discovering optimal programmatic policies that are highly interpretable. The code of this work is a

vailable at https://github.com/RU-Automated-Reasoning-Group/pi-PRL.
****

Mind Your Solver! On Adversarial Attack and Defense for Combinatorial Optimization

Han Lu,Zenan Li,Runzhong Wang,Qibing Ren,Junchi Yan,Zhigang Hua,Gan Liu,JUN ZHOU,Xiaokang Yang

Combinatorial optimization (CO) is a long-standing challenging task not only in its inherent complexity (e.g. NP-hard) but also the possible sensitivity to input conditions. In this paper, we take an initiative on developing the mechanisms for adversarial attack and defense towards combinatorial optimization solvers, whereby the solver is treated as a black-box function and the original problem's underlying graph structure (which is often available and associated with the problem instance, e.g. DAG, TSP) is attacked under a given budget. Experimental results on three real-world combinatorial optimization problems reveal the vulnerability of existing solvers to adversarial attack, including the commercial solvers like Gurobi. In particular, we present a simple yet effective defense strategy to modify the graph structure to increase the robustness of solvers, which shows its universal effectiveness across tasks and solvers.
****

Tactics on Refining Decision Boundary for Improving Certification-based Robust Training

Wang Zhang,Lam M. Nguyen,Subhro Das,Pin-Yu Chen,Sijia Liu,Alexandre Megretski,Luca Daniel,Tsui-Wei Weng

In verification-based robust training, existing methods utilize relaxation based methods to bound the worst case performance of neural networks given certain perturbation. However, these certification based methods treat all the examples equally regardless of their vulnerability and true adversarial distribution, limiting the model's potential in achieving optimal verifiable accuracy. In the paper, we propose new methods to include the customized weight distribution and automatic schedule tuning methods on the perturbation schedule. These methods are generally applicable to all the verification-based robust training with almost no additional computational cost. Our results show improvement on MNIST with $\epsilon = 0.3$ and CIFAR on $\epsilon = 8/255$ for both IBP and CROWN-IBP based methods.
****

Learning to Solve an Order Fulfillment Problem in Milliseconds with Edge-Feature-Embedded Graph Attention

Jingwei Yang,Qingchun Hou,Xiaoqing Wang,Yang Wei,Yuming Deng,Hongyang Jia,Ning Zhang

The order fulfillment problem is one of the fundamental combinatorial optimization problems in supply chain management and it is required to be solved in real-time for modern online retailing. Such a problem is computationally hard to address by exact mathematical programming methods. In this paper, we propose a machine learning method to solve it in milliseconds by formulating a tripartite graph and learning the best assignment policy through the proposed edge-feature-embedded graph attention mechanism. The edge-feature-embedded graph attention considers the high-dimensional edge features and accounts for the heterogeneous information, which are important characteristics of the studied optimization problem. The model is also size-invariant for problem instances of any scale, and it can address cases that are completely unseen during training. Experiments show that our model substantially outperforms the baseline heuristic method in optimality. The online inference time is milliseconds, which is thousands of times faster than the exact mathematical programming methods.
****

Active Hierarchical Exploration with Stable Subgoal Representation Learning

Siyuan Li,Jin Zhang,Jianhao Wang,Yang Yu,Chongjie Zhang

Goal-conditioned hierarchical reinforcement learning (GCHRL) provides a promising approach to solving long-horizon tasks. Recently, its success has been extended to more general settings by concurrently learning hierarchical policies and subgoal representations. Although GCHRL possesses superior exploration ability by decomposing tasks via subgoals, existing GCHRL methods struggle in temporally extended tasks with sparse external rewards, since the high-level policy learning relies on external rewards. As the high-level policy selects subgoals in an online learned representation space, the dynamic change of the subgoal space severely hinders effective high-level exploration. In this paper, we propose a novel regularization that contributes to both stable and efficient subgoal representation learning. Building upon the stable representation, we design measures of novelty and potential for subgoals, and develop an active hierarchical exploration strategy that seeks out new promising subgoals and states without intrinsic rewards. Experimental results show that our approach significantly outperforms state-of-the-art baselines in continuous control tasks with sparse rewards.

****

Learning Universal User Representations via Self-Supervised Lifelong Behaviors Modeling

Bei Yang,Ke Liu,Xiaoxiao Xu,Renjun Xu,Hong Liu,huan xu

Universal user representation is an important research topic in industry, and is widely used in diverse downstream user analysis tasks, such as user profiling and user preference prediction. With the rapid development of Internet service platforms, extremely long user behavior sequences have been accumulated. However, existing researches have little ability to model universal user representation based on lifelong behavior sequences since user registration. In this study, we propose a novel framework called Lifelong User Representation Model (LURM) to tackle this challenge. Specifically, LURM consists of two cascaded sub-models: (i) Bag of Interests (BoI) encodes user behaviors in any time period into a sparse vector with super-high dimension (eg. 10^5); (ii) Self-supervised Multi-anchor Encoder Network (SMEN) maps sequences of BoI features to multiple low-dimensional user representations by contrastive learning. SMEN achieves almost lossless dimensionality reduction with the main help of a novel multi-anchor module which can learn different aspects of user preferences. Experiments on several benchmark datasets show that our approach can outperform state-of-the-art unsupervised representation methods in downstream tasks.

****

DAIR: Disentangled Attention Intrinsic Regularization for Safe and Efficient Bimanual Manipulation

Minghao Zhang,Pingcheng Jian,Yi Wu,Huazhe Xu,Xiaolong Wang

We address the problem of safely solving complex bimanual robot manipulation tasks with sparse rewards. Such challenging tasks can be decomposed into sub-tasks that are accomplishable by different robots concurrently or sequentially for better efficiency. While previous reinforcement learning approaches primarily focus on modeling the compositionality of sub-tasks, two fundamental issues are largely ignored particularly when learning cooperative strategies for two robots: (i) domination, i.e., one robot may try to solve a task by itself and leaves the other idle; (ii) conflict, i.e., one robot can interrupt another's workspace when executing different sub-tasks simultaneously, which leads to unsafe collisions. To tackle these two issues, we propose a novel technique called disentangled attention, which provides an intrinsic regularization for two robots to focus on separate sub-tasks and objects. We evaluate our method on five bimanual manipulation tasks. Experimental results show that our proposed intrinsic regularization successfully avoids domination and reduces conflicts for the policies, which leads to significantly more efficient and safer cooperative strategies than all the baselines. Our project page with videos is at https://bimanual-attention.github.io/.

****

Deep AutoAugment

Yu Zheng,Zhi Zhang,Shen Yan,Mi Zhang

While recent automated data augmentation methods lead to state-of-the-art results, their design spaces and the derived data augmentation strategies still incorporate strong human priors. In this work, instead of fixing a set of hand-picked default augmentations alongside the searched data augmentations, we propose a fully automated approach for data augmentation search named Deep AutoAugment (Deep

AA). DeepAA progressively builds a multi-layer data augmentation pipeline from scratch by stacking augmentation layers one at a time until reaching convergence. For each augmentation layer, the policy is optimized to maximize the cosine similarity between the gradients of the original and augmented data along the direction with low variance. Our experiments show that even without default augmentations, we can learn an augmentation policy that achieves strong performance with that of previous works. Extensive ablation studies show that the regularized gradient matching is an effective search method for data augmentation policies. Our code is available at: https://github.com/MSU-MLSys-Lab/DeepAA .

****

Temporal Alignment Prediction for Supervised Representation Learning and Few-Shot Sequence Classification

Bing Su,Ji-Rong Wen

Explainable distances for sequence data depend on temporal alignment to tackle sequences with different lengths and local variances. Most sequence alignment methods infer the optimal alignment by solving an optimization problem under pre-defined feasible alignment constraints, which not only is time-consuming, but also makes end-to-end sequence learning intractable. In this paper, we propose a learnable sequence distance called Temporal Alignment Prediction (TAP). TAP employs a lightweight convolutional neural network to directly predict the optimal alignment between two sequences, so that only straightforward calculations are required and no optimization is involved in inference. TAP can be applied in different distance-based machine learning tasks. For supervised sequence representation learning, we show that TAP trained with various metric learning losses achieves completive performances with much faster inference speed. For few-shot action classification, we apply TAP as the distance measure in the metric learning-based episode-training paradigm. This simple strategy achieves comparable results with state-of-the-art few-shot action recognition methods.

****

DIVERSIFY to Generalize: Learning Generalized Representations for Time Series Classification

Wang Lu,Jindong Wang,Yiqiang Chen,Xinwei Sun

Time series classification is an important problem in real world. Due to its nonstationary property that the distribution changes over time, it remains challenging to build models for generalization to unseen distributions. In this paper, we propose to view the time series classification problem from the distribution perspective. We argue that the temporal complexity attributes to the unknown latent distributions within. To this end, we propose DIVERSIFY to learn generalized representations for time series classification. DIVERSIFY takes an iterative process: it first obtains the worst-case distribution scenario via adversarial training, then matches the distributions between all segments. We also present some theoretical insights. Extensive experiments on gesture recognition, speech commands recognition, and sensor-based human activity recognition demonstrate that DIVERSIFY significantly outperforms other baselines while effectively characterizing the latent distributions by qualitative and quantitative analysis.

****

Anti-Oversmoothing in Deep Vision Transformers via the Fourier Domain Analysis: From Theory to Practice

Peihao Wang,Wenqing Zheng,Tianlong Chen,Zhangyang Wang

Vision Transformer (ViT) has recently demonstrated promise in computer vision problems. However, unlike Convolutional Neural Networks (CNN), it is known that the performance of ViT saturates quickly with depth increasing, due to the observed attention collapse or patch uniformity. Despite a couple of empirical solutions, a rigorous framework studying on this scalability issue remains elusive. In this paper, we first establish a rigorous theory framework to analyze ViT features from the Fourier spectrum domain. We show that the self-attention mechanism inherently amounts to a low-pass filter, which indicates when ViT scales up its depth, excessive low-pass filtering will cause feature maps to only preserve their Direct-Current (DC) component. We then propose two straightforward yet effective techniques to mitigate the undesirable low-pass limitation. The first techniq ue, termed AttnScale, decomposes a self-attention block into low-pass and high-pass components, then rescales and combines these two filters to produce an all-pass self-attention matrix. The second technique, termed FeatScale, re-weights feature maps on separate frequency bands to amplify the high-frequency signals. Both techniques are efficient and hyperparameter-free, while effectively overcoming relevant ViT training artifacts such as attention collapse and patch uniformity. By seamlessly plugging in our techniques to multiple ViT variants, we demonstrate that they consistently help ViTs benefit from deeper architectures, bringing up to 1.1% performance gains "for free" (e.g., with little parameter overhead). We publicly release our codes and pre-trained models at https://github.com/VITA-Group/ViT-Anti-Oversmoothing.

****

Zero-shot detection of daily objects in YCB video dataset

Wanqing Xia

To let robots be able to manipulate objects, they have to sense the location of objects. With the development of visual data collecting and processing technology, robots are gradually evolving to localize objects in a greater field of view rather than being limited to a small space where the object could appear. To train such a robot vision system, pictures of all the objects need to be taken under various orientations and illumination. In the traditional manufacturing environment, this is applicable since objects involved in the production process does not change frequently. However, in the vision of smart manufacturing and high-mix-low-volume production, parts and products for robots to handle may change frequently. Thus, it is unrealistic to re-training the vision system for new products and tasks. Under this situation, we discovered the necessity to introduce a hot concept which is zero-shot object detection. Zero-shot object detection is a subset of unsupervised learning, and it aims to detect novel objects in the image with the knowledge learned from and only from seen objects. With zero-shot object detection algorithm, time can be greatly saved from collecting training data and training the vision system. Previous works focus on detecting objects in outdoor scenes, such as bikes, car, people, and dogs. The detection of daily objects is actually more challenging since the knowledge can be learned from each object is very limited. In this work, we explore the zero-shot detection of daily objects in indoor scenes since the objects’ size and environment are closely related to the manufacturing setup. The YCB Video Dataset is used in this work, which contains 21 objects in various categories. To the best of our knowledge, no previous work has explored zero-shot detection in this object size level and on this dataset.

****

Self-ensemble Adversarial Training for Improved Robustness

Hongjun Wang,Yisen Wang

Due to numerous breakthroughs in real-world applications brought by machine intelligence, deep neural networks (DNNs) are widely employed in critical applications. However, predictions of DNNs are easily manipulated with imperceptible adversarial perturbations, which impedes the further deployment of DNNs and may result in profound security and privacy implications. By incorporating adversarial samples into the training data pool, adversarial training is the strongest principled strategy against various adversarial attacks among all sorts of defense methods. Recent works mainly focus on developing new loss functions or regularizers, attempting to find the unique optimal point in the weight space. But none of them taps the potentials of classifiers obtained from standard adversarial training, especially states on the searching trajectory of training. In this work, we are dedicated to the weight states of models through the training process and devise a simple but powerful \emph{Self-Ensemble Adversarial Training} (SEAT) method for yielding a robust classifier by averaging weights of history models. This considerably improves the robustness of the target model against several well known adversarial attacks, even merely utilizing the naive cross-entropy loss to supervise. We also discuss the relationship between the ensemble of predictions from different adversarially trained models and the prediction of weight-ensembled models, as well as provide theoretical and empirical evidence that the propose d self-ensemble method provides a smoother loss landscape and better robustness than both individual models and the ensemble of predictions from different classifiers. We further analyze a subtle but fatal issue in the general settings for the self-ensemble model, which causes the deterioration of the weight-ensembled method in the late phases.

****

SCformer: Segment Correlation Transformer for Long Sequence Time Series Forecasting

Dazhao Du,Bing Su,Zhewei Wei

Long-term time series forecasting is widely used in real-world applications such as financial investment, electricity management and production planning. Recently, transformer-based models with strong sequence modeling ability have shown the potential in this task. However, most of these methods adopt point-wise dependencies discovery, whose complexity increases quadratically with the length of time series, which easily becomes intractable for long-term prediction. This paper proposes a new Transformer-based model called SCformer, which replaces the canonical self-attention with efficient segment correlation attention (SCAttention) mechanism. SCAttention divides time series into segments by the implicit series periodicity and utilizes correlations between segments to capture long short-term dependencies. Besides, we design a dual task that restores past series with the predicted future series to make SCformer more stable. Extensive experiments on several datasets in various fields demonstrate that our SCformer outperforms other Transformer-based methods and training with the additional dual task can enhance the generalization ability of the prediction model.

****

SANE: Specialization-Aware Neural Network Ensemble

Ziyue Li,Kan Ren,XINYANG JIANG,Mingzhe Han,Haipeng Zhang,Dongsheng Li

Real-world data is often generated by some complex distribution, which can be approximated by a composition of multiple simpler distributions. Thus, it is intuitive to divide the complex model learning into training several simpler models, each of which specializes in one simple distribution. Ensemble learning is one way to realize specialization, and has been widely used in practical machine learning scenarios. Many ensemble methods propose to increase diversity of base models, which could potentially result in model specialization. However, our studies show that without explicitly enforcing specification, pursuing diversity may not be enough to achieve satisfactory ensemble performance. In this paper, we propose SANE --- an end-to-end ensemble learning method that actively enforces model specification, where base models are trained to specialize in sub-regions of a latent space representing the simple distribution composition, and aggregated based on their specialties. Experiments in several prediction tasks on both image datasets and tabular datasets demonstrate the superior performance of our proposed method over state-of-the-art ensemble methods.

****

Stabilized Likelihood-based Imitation Learning via Denoising Continuous Normalizing Flow

Xin Zhang,Yanhua Li,Ziming Zhang,Christopher Brinton,Zhenming Liu,Zhi-Li Zhang,Hui Lu,Zhihong Tian

State-of-the-art imitation learning (IL) approaches, e.g, GAIL, apply adversarial training to minimize the discrepancy between expert and learner behaviors, which is prone to unstable training and mode collapse. In this work, we propose SLIL – Stabilized Likelihood-based Imitation Learning – a novel IL approach that directly maximizes the likelihood of observing the expert demonstrations. SLIL is a two-stage optimization framework, where in stage one the expert state distribution is estimated via a new method for denoising continuous normalizing flow, and in stage two the learner policy is trained to match both the expert’s policy and state distribution. Experimental evaluation of SLIL compared with several baselines in ten different physics-based control tasks reveals superior results in terms of learner policy performance, training stability, and mode distribution preservation.

****

Adaptive Speech Duration Modification using a Deep-Generative Framework
Ravi Shankar,Archana Venkataraman

We propose the first method to adaptively modify the duration of a given speechsignal. Our approach uses a Bayesian framework to define a latent attention mapthat links frames of the input and target utterances. We train a masked convolu-tional encoder-decoder network to generate this attention map via a stochastic ver-sion of the mean absolute error loss function. Our model also predicts the lengthof the target speech signal using the encoder embeddings, which determines thenumber of time steps for the decoding operation. During testing, we generate theattention map as a proxy for the similarity matrix between the given input speechand an unknown target speech signal. Using this similarity matrix, we compute awarping path of alignment between the two signals. Our experiments demonstratethat this adaptive framework produces similar results to dynamic time warping,which relies on a known target signal, on both voice conversion and emotion con-version tasks. We also show that the modified speech utterances achieve high userquality ratings, thus highlighting the practical utility of our method.

****

Do deep networks transfer invariances across classes?
Allan Zhou,Fahim Tajwar,Alexander Robey,Tom Knowles,George J. Pappas,Hamed Hassani,Chelsea Finn

In order to generalize well, classifiers must learn to be invariant to nuisance transformations that do not alter an input's class. Many problems have "class-agnostic" nuisance transformations that apply similarly to all classes, such as lighting and background changes for image classification. Neural networks can learn these invariances given sufficient data, but many real-world datasets are heavily class imbalanced and contain only a few examples for most of the classes. We therefore pose the question: how well do neural networks transfer class-agnostic invariances learned from the large classes to the small ones? Through careful experimentation, we observe that invariance to class-agnostic transformations is still heavily dependent on class size, with the networks being much less invariant on smaller classes. This result holds even when using data balancing techniques, and suggests poor invariance transfer across classes. Our results provide one explanation for why classifiers generalize poorly on unbalanced and long-tailed distributions. Based on this analysis, we show how a generative approach for learning the nuisance transformations can help transfer invariances across classes and improve performance on a set of imbalanced image classification benchmarks.

****

FED-χ^2: Secure Federated Correlation Test
Lun Wang,Qi Pang,Shuai Wang,Dawn Song

In this paper, we propose the first secure federated χ^2-test protocol, FED-χ^2. We recast χ^2-test as a problem of the second moment estimation and use stable projection to encode the local information in a short vector. Due to the fact that such encodings can be aggregated with summation, secure aggregation can smoothly be applied to conceal the individual updates. We formally establish the security guarantee of FED-χ^2 by demonstrating that the joint distribution is hidden in a subspace containing exponentially possible distributions. Our evaluation results show that FED-χ^2 achieves good accuracy with small client-side computation overhead. FED-χ^2 performs comparably to the centralized χ^2-test in several real-world case studies. The code for evaluation is in the supplementary material.

****

Cross-Trajectory Representation Learning for Zero-Shot Generalization in RL
Bogdan Mazoure,Ahmed M Ahmed,R Devon Hjelm,Andrey Kolobov,Patrick MacAlpine

A highly desirable property of a reinforcement learning (RL) agent -- and a major difficulty for deep RL approaches -- is the ability to generalize policies learned on a few tasks over a high-dimensional observation space to similar tasks not seen during training. Many promising approaches to this challenge consider RL as a process of training two functions simultaneously: a complex nonlinear encoder that maps high-dimensional observations to a latent representation space, an d a simple linear policy over this space. We posit that a superior encoder for zero-shot generalization in RL can be trained by using solely an auxiliary SSL objective if the training process encourages the encoder to map behaviorally similar observations to similar representations, as reward-based signal can cause overfitting in the encoder (Raileanu et al., 2021). We propose Cross-Trajectory Representation Learning (CTRL), a method that runs within an RL agent and conditions its encoder to recognize behavioral similarity in observations by applying a novel SSL objective to pairs of trajectories from the agent's policies. CTRL can be viewed as having the same effect as inducing a pseudo-bisimulation metric but, crucially, avoids the use of rewards and associated overfitting risks. Our experiments ablate various components of CTRL and demonstrate that in combination with PPO it achieves better generalization performance on the challenging Procgen benchmark suite (Cobbe et al., 2020).

****

Vision-Based Manipulators Need to Also See from Their Hands

Kyle Hsu,Moo Jin Kim,Rafael Rafailov,Jiajun Wu,Chelsea Finn

We study how the choice of visual perspective affects learning and generalization in the context of physical manipulation from raw sensor observations. Compared with the more commonly used global third-person perspective, a hand-centric (eye-in-hand) perspective affords reduced observability, but we find that it consistently improves training efficiency and out-of-distribution generalization. These benefits hold across a variety of learning algorithms, experimental settings, and distribution shifts, and for both simulated and real robot apparatuses. However, this is only the case when hand-centric observability is sufficient; otherwise, including a third-person perspective is necessary for learning, but also harms out-of-distribution generalization. To mitigate this, we propose to regularize the third-person information stream via a variational information bottleneck. On six representative manipulation tasks with varying hand-centric observability adapted from the Meta-World benchmark, this results in a state-of-the-art reinforcement learning agent operating from both perspectives improving its out-of-distribution generalization on every task. While some practitioners have long put cameras in the hands of robots, our work systematically analyzes the benefits of doing so and provides simple and broadly applicable insights for improving end-to-end learned vision-based robotic manipulation.

****

When Vision Transformers Outperform ResNets without Pre-training or Strong Data Augmentations

Xiangning Chen,Cho-Jui Hsieh,Boqing Gong

Vision Transformers (ViTs) and MLPs signal further efforts on replacing hand-wired features or inductive biases with general-purpose neural architectures. Existing works empower the models by massive data, such as large-scale pre-training and/or repeated strong data augmentations, and still report optimization-related problems (e.g., sensitivity to initialization and learning rates). Hence, this paper investigates ViTs and MLP-Mixers from the lens of loss geometry, intending to improve the models' data efficiency at training and generalization at inference. Visualization and Hessian reveal extremely sharp local minima of converged models. By promoting smoothness with a recently proposed sharpness-aware optimizer, we substantially improve the accuracy and robustness of ViTs and MLP-Mixers on various tasks spanning supervised, adversarial, contrastive, and transfer learning (e.g., +5.3\% and +11.0\% top-1 accuracy on ImageNet for ViT-B/16 and Mixer-B/16, respectively, with the simple Inception-style preprocessing). We show that the improved smoothness attributes to sparser active neurons in the first few layers. The resultant ViTs outperform ResNets of similar size and throughput when trained from scratch on ImageNet without large-scale pre-training or strong data augmentations. Model checkpoints are available at \url{https://github.com/google-research/vision_transformer}.

****

Meta-Learning with Fewer Tasks through Task Interpolation

Huaxiu Yao,Linjun Zhang,Chelsea Finn

Meta-learning enables algorithms to quickly learn a newly encountered task with just a few labeled examples by transferring previously learned knowledge. However, the bottleneck of current meta-learning algorithms is the requirement of a large number of meta-training tasks, which may not be accessible in real-world scenarios. To address the challenge that available tasks may not densely sample the space of tasks, we propose to augment the task set through interpolation. By meta-learning with task interpolation (MLTI), our approach effectively generates additional tasks by randomly sampling a pair of tasks and interpolating the corresponding features and labels. Under both gradient-based and metric-based meta-learning settings, our theoretical analysis shows MLTI corresponds to a data-adaptive meta-regularization and further improves the generalization. Empirically, in our experiments on eight datasets from diverse domains including image recognition, pose prediction, molecule property prediction, and medical image classification, we find that the proposed general MLTI framework is compatible with representative meta-learning algorithms and consistently outperforms other state-of-the-art strategies.

****

Image Functions In Neural Networks: A Perspective On Generalization

Arushi Gupta

In this work, we show that training with SGD on ReLU neural networks gives rise to a natural set of functions for each image that are not perfectly correlated until later in training. Furthermore, we show experimentally that the intersection of paths for different images also changes during the course of training. We hypothesize that this lack of correlation and changing intersection may be a factor in explaining generalization, because it encourages the model to use different features at different times, and pass the same image through different functions during training. This may improve generalization in two ways. 1) By encouraging the model to learn the same image in different ways, and learn different commonalities between images, comparable to model ensembling. 2) By improving algorithmic stability, as for a particular feature, the model is not always reliant on the same set of images, so the removal of an image may not adversely affect the loss.

****

Uncertainty-Aware Deep Video Compression with Ensembles

Wufei Ma,Jiahao Li,Bin Li,Yan Lu

Deep learning-based video compression is a challenging task and many previous state-of-the-art learning-based video codecs use optical flows to exploit the temporal correlation between successive frames and then compress the residual error. Although these two-stage models are end-to-end optimized, errors in the intermediate errors are propagated to later stages and would harm the overall performance. In this work, we investigate the inherent uncertainty in these intermediate predictions and present an ensemble-based video compression model to capture the predictive uncertainty. We also propose an ensemble-aware loss to encourage the diversity between ensemble members and investigate the benefit of incorporating adversarial training in the video compression task. Experimental results on 1080p sequences show that our model can effectively save bits by more than 20% compared to DVC Pro.

****

On the Convergence of Projected Alternating Maximization for Equitable and Optimal Transport

Minhui Huang,Shiqian Ma,Lifeng Lai

This paper studies the equitable and optimal transport (EOT) problem, which has many applications such as fair division problems and optimal transport with multiple agents etc. In the discrete distributions case, the EOT problem can be formulated as a linear program (LP). Since this LP is prohibitively large for general LP solvers, Scetbon \etal suggests to perturb the problem by adding an entropy regularization. They proposed a projected alternating maximization algorithm (PAM) to solve the dual of the entropy regularized EOT. In this paper, we provide the first convergence analysis of PAM. A novel rounding procedure is proposed to help construct the primal solution for the original EOT problem. We also propose a variant of PAM by incorporating the extrapolation technique that can numeric ally improve the performance of PAM. Results in this paper may shed lights on block coordinate (gradient) descent methods for general optimization problems.

****

Unified Recurrence Modeling for Video Action Anticipation

Tsung-Ming Tai,Giuseppe Fiameni,Cheng-Kuang Lee,Simon See,Oswald Lanz

Forecasting future events based on evidence of current conditions is an innate skill of human beings, and key for predicting the outcome of any decision making. In artificial vision for example, we would like to predict the next human action before it is actually performed, without observing the future video frames associated to it. Computer vision models for action anticipation are expected to collect the subtle evidence in the preamble of the target actions. In prior studies recurrence modeling often leads to better performance, and the strong temporal inference is assumed to be a key element for reasonable prediction. To this end, we propose a unified recurrence modeling for video action anticipation by generalizing the recurrence mechanism from sequence into graph representation via message passing. The information flow in space-time can be described by the interaction between vertices and edges, and the changes of vertices for each incoming frame reflects the underlying dynamics. Our model leverages self-attention for all building blocks in the graph modeling, and we introduce different edge learning strategies can be end-to-end optimized while updating the vertices. Our experimental results demonstrate that our modeling method is light-weight, efficient, and outperforms all previous works on the large-scale EPIC-Kitchen dataset.

****

Universal Controllers with Differentiable Physics for Online System Identification

Michelle Guo,Wenhao Yu,Daniel Ho,Jiajun Wu,Yunfei Bai,Karen Liu,Wenlong Lu

Creating robots that can handle changing or unknown environments is a critical step towards real-world robot applications. Existing methods tackle this problem by training controllers robust to large ranges of environment parameters (Domain Randomization), or by combining ``Universal'' Controllers (UC) conditioned on environment parameters with learned identification modules that (implicitly or explicitly) identify the environment parameters from sensory inputs (Domain Adaptation). However, these methods can lead to over-conservative behaviors or poor generalization outside the training distribution. In this work, we present a domain adaptation approach that improves generalization of the identification module by leveraging prior knowledge in physics. Our proposed algorithm, UC-DiffOSI, combines a UC trained on a wide range of environments with an Online System Identification module based on a differentiable physics engine (DiffOSI). We evaluate UC-DiffOSI on articulated rigid body control tasks, including a wiping task that requires contact-rich environment interaction.

Compared to previous works, UC-DiffOSI outperforms domain randomization baselines and is more robust than domain adaptation methods that rely on learned identification models. In addition, we perform two studies showing that UC-DiffOSI operates well in environments with changing or unknown dynamics. These studies test sudden changes in the robot's mass and inertia, and they evaluate in an environment (PyBullet) whose dynamics differs from training (NimblePhysics).

****

Identifying the Limits of Cross-Domain Knowledge Transfer for Pretrained Models

Zhengxuan Wu,Nelson F. Liu,Christopher Potts

There is growing evidence that pretrained language models improve task-specific fine-tuning even where the task examples are radically different from those seen in training. What is the nature of this surprising cross-domain transfer? We offer a partial answer via a systematic exploration of how much transfer occurs when models are denied any information about word identity via random scrambling. In four classification tasks and two sequence labeling tasks, we evaluate LSTMs using GloVe embeddings, BERT, and baseline models. Among these models, we find that only BERT shows high rates of transfer into our scrambled domains, and for classification but not sequence labeling tasks. Our analyses seek to explain why transfer succeeds for some tasks but not others, to isolate the separate contributions of pretraining versus fine-tuning, to show that the fine-tuning process i

s not merely learning to unscramble the scrambled inputs, and to quantify the role of word frequency. These findings help explain where and why cross-domain transfer occurs, which can guide future studies and practical fine-tuning efforts.
****

On Covariate Shift of Latent Confounders in Imitation and Reinforcement Learning
Guy Tennenholtz,Assaf Hallak,Gal Dalal,Shie Mannor,Gal Chechik,Uri Shalit
We consider the problem of using expert data with unobserved confounders for imitation and reinforcement learning. We begin by defining the problem of learning from confounded expert data in a contextual MDP setup. We analyze the limitations of learning from such data with and without external reward and propose an adjustment of standard imitation learning algorithms to fit this setup. In addition, we discuss the problem of distribution shift between the expert data and the online environment when partial observability is present in the data. We prove possibility and impossibility results for imitation learning under arbitrary distribution shift of the missing covariates. When additional external reward is provided, we propose a sampling procedure that addresses the unknown shift and prove convergence to an optimal solution. Finally, we validate our claims empirically on challenging assistive healthcare and recommender system simulation tasks.
****

Multi-Agent Constrained Policy Optimisation
Shangding Gu,Jakub Grudzien Kuba,Muning Wen,Ruiqing Chen,Ziyan Wang,Zheng Tian,Jun Wang,Alois Knoll,Yaodong Yang
Developing reinforcement learning algorithms that satisfy safety constraints is becoming increasingly important in real-world applications. In multi-agent reinforcement learning (MARL) settings, policy optimisation with safety awareness is particularly challenging because each individual agent has to not only meet its own safety constraints, but also consider those of others so that their joint behaviour can be guaranteed safe. Despite its importance, the problem of safe multi-agent learning has not been rigorously studied; very few solutions have been proposed, nor a sharable testing environment or benchmarks. To fill these gaps, in this work, we formulate the safe MARL problem as a constrained Markov game and solve it with policy optimisation methods. Our solutions---Multi-Agent Constrained Policy Optimisation (MACPO) and MAPPO-Lagrangian---leverage the theories from both constrained policy optimisation and multi-agent trust region learning. Crucially, our methods enjoy theoretical guarantees of both monotonic improvement in reward and satisfaction of safety constraints at every iteration. To examine the effectiveness of our methods, we develop the benchmark suite of Safe Multi-Agent MuJoCo that involves a variety of MARL baselines. Experimental results justify that MACPO/MAPPO-Lagrangian can consistently satisfy safety constraints, meanwhile achieving comparable performance to strong baselines.
****

RvS: What is Essential for Offline RL via Supervised Learning?
Scott Emmons,Benjamin Eysenbach,Ilya Kostrikov,Sergey Levine
Recent work has shown that supervised learning alone, without temporal difference (TD) learning, can be remarkably effective for offline RL. When does this hold true, and which algorithmic components are necessary? Through extensive experiments, we boil supervised learning for offline RL down to its essential elements. In every environment suite we consider, simply maximizing likelihood with a two-layer feedforward MLP is competitive with state-of-the-art results of substantially more complex methods based on TD learning or sequence modeling with Transformers. Carefully choosing model capacity (e.g., via regularization or architecture) and choosing which information to condition on (e.g., goals or rewards) are critical for performance. These insights serve as a field guide for practitioners doing Reinforcement Learning via Supervised Learning (which we coin RvS learning). They also probe the limits of existing RvS methods, which are comparatively weak on random data, and suggest a number of open problems.
****

Differentiable Self-Adaptive Learning Rate
Bozhou Chen,Hongzhi Wang,Chenmin Ba
Adaptive learning rate has been studied for a long time. In the training session

of neural networks, learning rate controls update stride and direction in a multi-dimensional space. A large learning rate may cause failure to converge, while a small learning rate will make the convergence too slow.

Even though some optimizers make learning rate adaptive to the training, e.g., using first-order and second-order momentum to adapt learning rate, their network's parameters are still unstable during training and converges too slowly in many occasions.

To solve this problem, we propose a novel optimizer which makes learning rate differentiable with the goal of minimizing loss function and thereby realize an optimizer with truly self-adaptive learning rate. We conducted extensive experiments on multiple network models compared with various benchmark optimizers. It is shown that our optimizer achieves fast and high qualified convergence in extremely short epochs, which is far more faster than those state-of-art optimizers.

****

Iterative Hierarchical Attention for Answering Complex Questions over Long Documents

Haitian Sun,William W. Cohen,Ruslan Salakhutdinov

We propose a new model, DocHopper, that iteratively attends to different parts of long, hierarchically structured documents to answer complex questions. Similar to multi-hop question-answering (QA) systems, at each step, DocHopper uses a query q to attend to information from a document, combines this “retrieved” information with q to produce the next query. However, in contrast to most previous multi-hop QA systems, DocHopper is able to “retrieve” either short passages or long sections of the document, thus emulating a multi-step process of “navigating” through a long document to answer a question. To enable this novel behavior, DocHopper does not combine document information with q by concatenating text to the text of q, but by combining a compact neural representation of q with a compact neural representation of a hierarchical part of the document -- potentially a large part. We experiment with DocHopper on four different QA tasks that require reading long and complex documents to answer multi-hop questions, and show that DocHopper outperforms all baseline models and achieves state-of-the-art results on all datasets. Additionally, DocHopper is efficient at inference time, being 3 - 10 times faster than the baselines.

****

Differentiable Hyper-parameter Optimization

Bozhou Chen,Hongzhi Wang,Chenmin Ba

Hyper-parameters are widely present in machine learning.

Concretely, large amount of hyper-parameters exist in network layers, such as kernel size, channel size and the hidden layer size, which directly affect performance of the model.

Thus, hyper-parameter optimization is crucial for machine learning. Current hyper-parameter optimization always requires multiple training sessions, resulting in a large time consuming.

To solve this problem, we propose a method to fine-tune neural network's hyper-parameters efficiently in this paper, where optimization completes in only one training session.

We apply our method for the optimization of various neural network layers' hyper-parameters and compare it with multiple benchmark hyper-parameter optimization models.

Experimental results show that our method is commonly 10 times faster than traditional and mainstream methods such as random search, Bayesian optimization and many other state-of-art models. It also achieves higher quality hyper-parameters with better accuracy and stronger stability.

****

Learning Rich Nearest Neighbor Representations from Self-supervised Ensembles

Bram Wallace,Devansh Arpit,Huan Wang,Caiming Xiong

Pretraining convolutional neural networks via self-supervision, and applying them in transfer learning, is an incredibly fast-growing field that is rapidly and iteratively improving performance across practically all image domains.

Meanwhile, model ensembling is one of the most universally applicable techniques

in supervised learning literature and practice, offering a simple solution to reliably improve performance. But how to optimally combine self-supervised models to maximize representation quality has largely remained unaddressed.
In this work, we provide a framework to perform self-supervised model ensembling via a novel method of learning representations directly through gradient descent at inference time.
This technique improves representation quality, as measured by k-nearest neighbors, both on the in-domain dataset and in the transfer setting, with models transferable from the former setting to the latter.
Additionally, this direct learning of feature through backpropagation improves representations from even a single model, echoing the improvements found in self-distillation.
****

Online Hyperparameter Meta-Learning with Hypergradient Distillation

Hae Beom Lee,Hayeon Lee,JaeWoong Shin,Eunho Yang,Timothy Hospedales,Sung Ju Hwang

Many gradient-based meta-learning methods assume a set of parameters that do not participate in inner-optimization, which can be considered as hyperparameters. Although such hyperparameters can be optimized using the existing gradient-based hyperparameter optimization (HO) methods, they suffer from the following issues. Unrolled differentiation methods do not scale well to high-dimensional hyperparameters or horizon length, Implicit Function Theorem (IFT) based methods are restrictive for online optimization, and short horizon approximations suffer from short horizon bias. In this work, we propose a novel HO method that can overcome these limitations, by approximating the second-order term with knowledge distillation. Specifically, we parameterize a single Jacobian-vector product (JVP) for each HO step and minimize the distance from the true second-order term. Our method allows online optimization and also is scalable to the hyperparameter dimension and the horizon length. We demonstrate the effectiveness of our method on three different meta-learning methods and two benchmark datasets.
****

NeuroSED: Learning Subgraph Similarity via Graph Neural Networks

Rishabh Ranjan,Siddharth Grover,Sourav Medya,Venkatesan Chakaravarthy,Yogish Sabharwal,Sayan Ranu

Subgraph similarity search is a fundamental operator in graph analysis. In this framework, given a query graph and a graph database, the goal is to identify subgraphs of the database graphs that are structurally similar to the query. Subgraph edit distance (SED) is one of the most expressive measures of subgraph similarity. In this work, we study the problem of learning SED from a training set of graph pairs and their SED values. Towards that end, we design a novel siamese graph neural network called NeuroSED, which learns an embedding space with a rich structure reminiscent of SED. With the help of a specially crafted inductive bias, NeuroSED not only enables high accuracy but also ensures that the predicted SED, like true SED, satisfies triangle inequality. The design is generic enough to also model graph edit distance (GED), while ensuring that the predicted GED space is metric, like the true GED space. Extensive experiments on real graph datasets, for both SED and GED, establish that NeuroSED achieves ≈ 2 times lower RMSE than the state of the art and is ≈ 18 times faster than the fastest baseline. Further, owing to its pair-independent embeddings and theoretical properties, NeuroSED allows orders-of-magnitude faster graph/subgraph retrieval.
****

Auto-Encoding Inverse Reinforcement Learning

Kaifeng Zhang,Rui Zhao,Ziming Zhang,Yang Gao

Reinforcement learning (RL) provides a powerful framework for decision-making, but its application in practice often requires a carefully designed reward function. Inverse Reinforcement Learning (IRL) has shed light on automatic reward acquisition, but it is still difficult to apply IRL to solve real-world tasks. In this work, we propose Auto-Encoding Inverse Reinforcement Learning (AEIRL), a robust and scalable IRL framework, which belongs to the adversarial imitation learni ng class. To recover reward functions from expert demonstrations, AEIRL utilizes the reconstruction error of an auto-encoder as the learning signal, which provides more information for optimizing policies, compared to the binary logistic loss. Subsequently, we use the derived objective functions to train the reward function and the RL agent. Experiments show that AEIRL performs superior in comparison with state-of-the-art methods in the MuJoCo environments. More importantly, in more realistic settings, AEIRL shows much better robustness when the expert demonstrations are noisy. Specifically, our method achieves 16% relative improvement compared to the best baseline FAIRL on clean expert data and 38% relative improvement compared to the best baseline PWIL on noisy expert data both with the metric overall averaged scaled rewards.

****

LEARNING GUARANTEES FOR GRAPH CONVOLUTIONAL NETWORKS ON THE STOCHASTIC BLOCK MODEL

Wei Lu

An abundance of neural network models and algorithms for diverse tasks on graphs have been developed in the past five years. However, very few provable guarantees have been available for the performance of graph neural network models. This state of affairs is in contrast with the steady progress on the theoretical underpinnings of traditional dense and convolutional neural networks. In this paper we present the first provable guarantees for one of the best-studied families of graph neural network models, Graph Convolutional Networks (GCNs), for semi- supervised community detection tasks. We show that with high probability over the initialization and training data, a GCN will efficiently learn to detect communities on graphs drawn from a stochastic block model. Our proof relies on a fine-grained analysis of the training dynamics in order to overcome the complexity of a non-convex optimization landscape with many poorly-performing local minima.

****

Learning Versatile Neural Architectures by Propagating Network Codes

Mingyu Ding,Yuqi Huo,Haoyu Lu,Linjie Yang,Zhe Wang,Zhiwu Lu,Jingdong Wang,Ping Luo

This work explores how to design a single neural network capable of adapting to multiple heterogeneous vision tasks, such as image segmentation, 3D detection, and video recognition. This goal is challenging because both network architecture search (NAS) spaces and methods in different tasks are inconsistent. We solve this challenge from both sides. We first introduce a unified design space for multiple tasks and build a multitask NAS benchmark (NAS-Bench-MR) on many widely used datasets, including ImageNet, Cityscapes, KITTI, and HMDB51. We further propose Network Coding Propagation (NCP), which back-propagates gradients of neural predictors to directly update architecture codes along the desired gradient directions to solve various tasks. In this way, optimal architecture configurations can be found by NCP in our large search space in seconds.

Unlike prior arts of NAS that typically focus on a single task, NCP has several unique benefits. (1) NCP transforms architecture optimization from data-driven to architecture-driven, enabling joint search an architecture among multitasks with different data distributions. (2) NCP learns from network codes but not original data, enabling it to update the architecture efficiently across datasets. (3) In addition to our NAS-Bench-MR, NCP performs well on other NAS benchmarks, such as NAS-Bench-201. (4) Thorough studies of NCP on inter-, cross-, and intra-tasks highlight the importance of cross-task neural architecture design, i.e., multitask neural architectures and architecture transferring between different tasks. Code is available at https://github.com/dingmyu/NCP.

****

Fight fire with fire: countering bad shortcuts in imitation learning with good shortcuts

Chuan Wen,Jianing Qian,Jierui Lin,Dinesh Jayaraman,Yang Gao

When operating in partially observed settings, it is important for a control policy to fuse information from a history of observations. However, a naive implementation of this approach has been observed repeatedly to fail for imitation-lea rned policies, often in surprising ways, and to the point of sometimes performing worse than when using instantaneous observations alone. We observe that behavioral cloning policies acting on single observations and observation histories each have their strengths and drawbacks, and combining them optimally could achieve the best of both worlds. Motivated by this, we propose a simple model combination approach inspired by human decision making: we first compute a coarse action based on the instantaneous observation, and then refine it into a final action using historical information. Our experiments show that this outperforms all baselines on CARLA autonomous driving from images and various MuJoCo continuous control tasks.

****
Tighter Sparse Approximation Bounds for ReLU Neural Networks
Carles Domingo-Enrich,Youssef Mroueh
A well-known line of work (Barron, 1993; Breiman, 1993; Klusowski & Barron, 2018) provides bounds on the width n of a ReLU two-layer neural network needed to approximate a function f over the ball $\mathcal{B}_R(\mathbb{R}^d)$ up to error ϵ, when the Fourier based quantity $C_f = \int_{\mathbb{R}^d} \|\xi\|^2 |\hat{f}(\xi)| \ d\xi$ is finite. More recently Ongie et al. (2019) used the Radon transform as a tool for analysis of infinite-width ReLU two-layer networks. In particular, they introduce the concept of Radon-based $\mathcal{R}$-norms and show that a function defined on $\mathbb{R}^d$ can be represented as an infinite-width two-layer neural network if and only if its $\mathcal{R}$-norm is finite. In this work, we extend the framework of Ongie et al. (2019) and define similar Radon-based semi-norms ($\mathcal{R}, \mathcal{U}$-norms) such that a function admits an infinite-width neural network representation on a bounded open set $\mathcal{U} \subseteq \mathbb{R}^d$ when its $\mathcal{R}, \mathcal{U}$-norm is finite. Building on this, we derive sparse (finite-width) neural network approximation bounds that refine those of Breiman (1993); Klusowski & Barron (2018). Finally, we show that infinite-width neural network representations on bounded open sets are not unique and study their structure, providing a functional view of mode connectivity.
****
Task-Induced Representation Learning
Jun Yamada,Karl Pertsch,Anisha Gunjal,Joseph J Lim
In this work, we evaluate the effectiveness of representation learning approaches for decision making in visually complex environments. Representation learning is essential for effective reinforcement learning (RL) from high-dimensional inputs. Unsupervised representation learning approaches based on reconstruction, prediction or contrastive learning have shown substantial learning efficiency gains. Yet, they have mostly been evaluated in clean laboratory or simulated settings. In contrast, real environments are visually complex and contain substantial amounts of clutter and distractors. Unsupervised representations will learn to model such distractors, potentially impairing the agent’s learning efficiency. In contrast, an alternative class of approaches, which we call task-induced representation learning, leverages task information such as rewards or demonstrations from prior tasks to focus on task-relevant parts of the scene and ignore distractors. We investi- gate the effectiveness of unsupervised and task-induced representation learning approaches on four visually complex environments, from Distracting DMControl to the CARLA driving simulator. For both, RL and imitation learning, we find that representation learning generally improves sample efficiency on unseen tasks even in visually complex scenes and that task-induced representations can double learning efficiency compared to unsupervised alternatives.
****
Long Expressive Memory for Sequence Modeling
T. Konstantin Rusch,Siddhartha Mishra,N. Benjamin Erichson,Michael W. Mahoney
We propose a novel method called Long Expressive Memory (LEM) for learning long-term sequential dependencies. LEM is gradient-based, it can efficiently process sequential tasks with very long-term dependencies, and it is sufficiently expressive to be able to learn complicated input-output maps. To derive LEM, we consid

er a system of multiscale ordinary differential equations, as well as a suitable time-discretization of this system. For LEM, we derive rigorous bounds to show the mitigation of the exploding and vanishing gradients problem, a well-known challenge for gradient-based recurrent sequential learning methods. We also prove that LEM can approximate a large class of dynamical systems to high accuracy. Our empirical results, ranging from image and time-series classification through dynamical systems prediction to speech recognition and language modeling, demonstrate that LEM outperforms state-of-the-art recurrent neural networks, gated recurrent units, and long short-term memory models.
****

In defense of dual-encoders for neural ranking

Aditya Krishna Menon,Sadeep Jayasumana,Seungyeon Kim,Ankit Singh Rawat,Sashank J. Reddi,Sanjiv Kumar

Transformer-based models such as BERT have proven successful in information retrieval problem, which seek to identify relevant documents for a given query. There are two broad flavours of such models: cross-attention (CA) models, which learn a joint embedding for the query and document, and dual-encoder (DE) models, which learn separate embeddings for the query and document. Empirically, CA models are often found to be more accurate, which has motivated a series of works seeking to bridge this gap. However, a more fundamental question remains less explored: does this performance gap reflect an inherent limitation in the capacity of DE models, or a limitation in the training of such models? And does such an understanding suggest a principled means of improving DE models? In this paper, we study these questions, with three contributions. First, we establish theoretically that with a sufficiently large embedding dimension, DE models have the capacity to model a broad class of score distributions. Second, we show empirically that on real-world problems, DE models may overfit to spurious correlations in the training set, and thus under-perform on test samples. To mitigate this behaviour, we propose a suitable distillation strategy, and confirm its practical efficacy on the MSMARCO-Passage and Natural Questions benchmarks.
****

Learning an Object-Based Memory System

Yilun Du,Joshua B. Tenenbaum,Tomás Lozano-Pérez,Leslie Pack Kaelbling

A robot operating in a household makes observations of multiple objects as it moves around over the course of days or weeks. The objects may be moved by inhabitants, but not completely at random. The robot may be called upon later to retrieve objects and will need a long-term object-based memory in order to know how to find them. In this paper, we combine some aspects of classic techniques for data-association filtering with modern attention-based neural networks to construct object-based memory systems that consume and produce high-dimensional observations and hypotheses. We perform end-to-end learning on labeled observation trajectories to learn both the internal transition and observation models. We demonstrate the system's effectiveness on a sequence of problem classes of increasing difficulty and show that it outperforms clustering-based methods, classic filters, and unstructured neural approaches.
****

Graph-based Nearest Neighbor Search in Hyperbolic Spaces

Liudmila Prokhorenkova,Dmitry Baranchuk,Nikolay Bogachev,Yury Demidovich,Alexander Kolpakov

The nearest neighbor search (NNS) problem is widely studied in Euclidean space, and graph-based algorithms are known to outperform other approaches for this task. However, hyperbolic geometry often allows for better data representation in various domains, including graphs, words, and images. In this paper, we show that graph-based approaches are also well suited for hyperbolic geometry. From a theoretical perspective, we rigorously analyze the time and space complexity of graph-based NNS, assuming that an n-element dataset is uniformly distributed within a d-dimensional ball of radius R in the hyperbolic space of curvature -1. Under some conditions on R and d, we derive the time and space complexity of graph-based NNS and compare the obtained results with known guarantees for the Euclidean case. Interestingly, in the dense setting ($d \ll \log n$) and unde r some assumptions on the radius R, graph-based NNS has lower time complexity in the hyperbolic space. This agrees with our experiments: we consider datasets embedded in hyperbolic and Euclidean spaces and show that graph-based NNS can be more efficient in the hyperbolic space. We also demonstrate that graph-based methods outperform other existing baselines for hyperbolic NNS. Overall, our theoretical and empirical analysis suggests that graph-based NNS can be considered a default approach for similarity search in hyperbolic spaces.

****

Model-Agnostic Meta-Attack: Towards Reliable Evaluation of Adversarial Robustness

Xiao Yang,Yinpeng Dong,Wenzhao Xiang,Tianyu Pang,Hang Su,Jun Zhu

The vulnerability of deep neural networks to adversarial examples has motivated an increasing number of defense strategies for promoting model robustness. However, the progress is usually hampered by insufficient robustness evaluations. As the de facto standard to evaluate adversarial robustness, adversarial attacks typically solve an optimization problem of crafting adversarial examples with an iterative process. In this work, we propose a Model-Agnostic Meta-Attack (MAMA) approach to discover stronger attack algorithms automatically. Our method learns the optimizer in adversarial attacks parameterized by a recurrent neural network, which is trained over a class of data samples and defenses to produce effective update directions during adversarial example generation. Furthermore, we develop a model-agnostic training algorithm to improve the generalization ability of the learned optimizer when attacking unseen defenses. Our approach can be flexibly incorporated with various attacks and consistently improves the performance with little extra computational cost. Extensive experiments demonstrate the effectiveness of the learned attacks by MAMA compared to the state-of-the-art attacks on different defenses, leading to a more reliable evaluation of adversarial robustness.

****

Polyphonic Music Composition: An Adversarial Inverse Reinforcement Learning Approach

Kelvin Xavier Munguia Velez,Von-Wun Soo

Most recent approaches to automatic music harmony composition adopt deep supervised learning to train a model using a set of human composed songs as training data. However, these approaches suffer from inherent limitations from the chosen deep learning models which may lead to unpleasing harmonies. This paper explores an alternative approach to harmony composition using a combination of novel Deep Supervised Learning, DeepReinforcement Learning and Inverse Reinforcement Learning techniques. In this novel approach, our model selects the next chord in the composition(action) based on the previous notes(states), therefore allowing us to model harmony composition as a reinforcement learning problem in which we look to maximize an overall accumulated reward. However, designing an appropriate reward function is known to be a very tricky and difficult process. To overcome this problem we propose learning a reward function from a set of human-composed tracks using Adversarial Inverse Reinforcement Learning. We start by training a Bi-axial LSTM model using supervised learning and improve upon it by tuning it using Deep Q-learning. Instead of using GANs to generate a similar music composition to human compositions directly, we adopt GANs to learn the reward function of the music trajectories from human compositions. We then combine the learned reward function with a reward based on music theory rules to improve the generation of the model trained by supervised learning. The results show improvement over a pre-trained model without training with reinforcement learning with respect to a set of objective metrics and preference from subjective user evaluation.

****

Classify and Generate Reciprocally: Simultaneous Positive-Unlabelled Learning and Conditional Generation with Extra Data

Bing Yu,Ke Sun,He Wang,Zhanxing Zhu,Zhouchen Lin

The scarcity of class-labeled data is a ubiquitous bottleneck in a wide range of machine learning problems. While abundant unlabeled data normally exist and provide a potential solution, it is extremely challenging to exploit them. In this paper, we address this problem by leveraging Positive-Unlabeled~(PU) classification and the conditional generation with extra unlabeled data \emph{simultaneously}, both of which aim to make full use of agnostic unlabeled data to improve classification and generation performance. In particular, we present a novel training framework to jointly target both PU classification and conditional generation when exposing to extra data, especially out-of-distribution unlabeled data, by exploring the interplay between them: 1) enhancing the performance of PU classifiers with the assistance of a novel Conditional Generative Adversarial Network~(CGAN) that is robust to noisy labels, 2) leveraging extra data with predicted labels from a PU classifier to help the generation. Our key contribution is a Classifier-Noise-Invariant Conditional GAN~(CNI-CGAN) that can learn the clean data distribution from noisy labels predicted by a PU classifier. Theoretically, we proved the optimal condition of CNI-CGAN and experimentally, we conducted extensive evaluations on diverse datasets, verifying the simultaneous improvements on both classification and generation.

****

DISCOVERING AND EXPLAINING THE REPRESENTATION BOTTLENECK OF DNNS

Huiqi Deng,Qihan Ren,Hao Zhang,Quanshi Zhang

This paper explores the bottleneck of feature representations of deep neural networks (DNNs), from the perspective of the complexity of interactions between input variables encoded in DNNs. To this end, we focus on the multi-order interaction between input variables, where the order represents the complexity of interactions. We discover that a DNN is more likely to encode both too simple and too complex interactions, but usually fails to learn interactions of intermediate complexity. Such a phenomenon is widely shared by different DNNs for different tasks. This phenomenon indicates a cognition gap between DNNs and humans, and we call it a representation bottleneck. We theoretically prove the underlying reason for the representation bottleneck. Furthermore, we propose losses to encourage/penalize the learning of interactions of specific complexities, and analyze the representation capacities of interactions of different complexities. The code is available at https://github.com/Nebularaid2000/bottleneck.

****

ImpressLearn: Continual Learning via Combined Task Impressions

Dhrupad Bhardwaj,Julia Kempe,Artem M Vysogorets,Angela Teng,Evaristus Ezekwem

This work proposes a new method to sequentially train a deep neural network on multiple tasks without suffering catastrophic forgetting, while endowing it with the capability to quickly adapt to unknown tasks. Starting from existing work on network masking (Wortsman et al., 2020), we show that a simple to learn linear combination of a small number of task-specific masks (”impressions”) ona randomly initialized backbone network is sufficient to both retain accuracy on previously learned tasks, as well as achieve high accuracy on new tasks.

In contrast to previous methods, we do not require to generate dedicated masks or contexts for each new task, instead leveraging transfer learning to keep per-task parameter overhead negligible. Our work illustrates the power of linearly combining individual impressions, each of which fares poorly in isolation, to achieve performance comparable to a dedicated mask. Moreover, even repeated impressions from the same task (homogeneous masks), when combined can approach the performance of heterogeneous combinations if sufficiently many impressions are used.

Our approach scales more efficiently than existing methods, requiring orders of magnitude fewer parameters and can function without modification even when task identity is missing. In addition, in the setting where task labels are not given at inference, our algorithm gives an often favorable alternative to the entropy based task-inference methods proposed in (Wortsman et al., 2020). We evaluate our method on a number of well known image classification data sets and architectures

****

Generative Models as a Data Source for Multiview Representation Learning
Ali Jahanian,Xavier Puig,Yonglong Tian,Phillip Isola

Generative models are now capable of producing highly realistic images that look nearly indistinguishable from the data on which they are trained. This raises the question: if we have good enough generative models, do we still need datasets? We investigate this question in the setting of learning general-purpose visual representations from a black-box generative model rather than directly from data. Given an off-the-shelf image generator without any access to its training data, we train representations from the samples output by this generator. We compare several representation learning methods that can be applied to this setting, using the latent space of the generator to generate multiple "views" of the same semantic content. We show that for contrastive methods, this multiview data can naturally be used to identify positive pairs (nearby in latent space) and negative pairs (far apart in latent space). We find that the resulting representations rival or even outperform those learned directly from real data, but that good performance requires care in the sampling strategy applied and the training method. Generative models can be viewed as a compressed and organized copy of a dataset, and we envision a future where more and more "model zoos" proliferate while datasets become increasingly unwieldy, missing, or private. This paper suggests several techniques for dealing with visual representation learning in such a future. Code is available on our project page https://ali-design.github.io/GenRep/.

****

Towards Generative Latent Variable Models for Speech
Jakob Drachmann Havtorn,Lasse Borgholt,Jes Frellsen,Søren Hauberg,Lars Maaløe

While stochastic latent variable models (LVMs) now achieve state-of-the-art performance on natural image generation, they are still inferior to deterministic models on speech. On natural images, these models have been parameterised with very deep hierarchies of latent variables, but research shows that these model constructs are not directly applicable to sequence data. In this paper, we benchmark popular temporal LVMs against state-of-the-art deterministic models on speech. We report the likelihood, which is a much used metric in the image domain but rarely, and often incomparably, reported for speech models. This is prerequisite work needed for the research community to improve LVMs on speech. We adapt Clockwork VAE, a state-of-the-art temporal LVM for video generation, to the speech domain, similar to how WaveNet adapted PixelCNN from images to speech. Despite being autoregressive only in latent space, we find that the Clockwork VAE outperforms previous LVMs and reduces the gap to deterministic models by using a hierarchy of latent variables.

****

Clustered Task-Aware Meta-Learning by Learning from Learning Paths
Danni Peng,Sinno Pan

To enable effective learning of new tasks with only few samples, meta-learning acquires common knowledge from the existing tasks with a globally shared meta-learner. To further address the problem of task heterogeneity, recent developments balance between customization and generalization by incorporating task clustering to generate the task-aware modulation to be applied on the global meta-learner. However, these methods learn task representation mostly from the features of input data, while the task-specific optimization process with respect to the base-learner model is often neglected. In this work, we propose a Clustered Task-Aware Meta-Learning (CTML) framework with task representation learned from its own learning path. We first conduct a rehearsed task learning from the common initialization, and collect a set of geometric quantities that adequately describes this learning path. By inputting this set of values into a meta path learner, we automatically abstract path representation optimized for the downstream clustering and modulation. To further save the computational cost incurred by the additional rehearsed learning, we devise a shortcut tunnel to directly map between the path and feature cluster assignments. Extensive experiments on two real-world application domains: few-shot image classification and cold-start recommendation demonstrate the superiority of CTML compared to state-of-the-art baselines.

****

GiraffeDet: A Heavy-Neck Paradigm for Object Detection

yiqi jiang,Zhiyu Tan,Junyan Wang,Xiuyu Sun,Ming Lin,Hao Li

In conventional object detection frameworks, a backbone body inherited from image recognition models extracts deep latent features and then a neck module fuses these latent features to capture information at different scales. As the resolution in object detection is much larger than in image recognition, the computational cost of the backbone often dominates the total inference cost. This heavy-backbone design paradigm is mostly due to the historical legacy when transferring image recognition models to object detection rather than an end-to-end optimized design for object detection. In this work, we show that such paradigm indeed leads to sub-optimal object detection models. To this end, we propose a novel heavy-neck paradigm, GiraffeDet, a giraffe-like network for efficient object detection. The GiraffeDet uses an extremely lightweight backbone and a very deep and large neck module which encourages dense information exchange among different spatial scales as well as different levels of latent semantics simultaneously. This design paradigm allows detectors to process the high-level semantic information and low-level spatial information at the same priority even in the early stage of the network, making it more effective in detection tasks. Numerical evaluations on multiple popular object detection benchmarks show that GiraffeDet consistently outperforms previous SOTA models across a wide spectrum of resource constraints. The source code is available at
https://github.com/jyqi/GiraffeDet.

****

Model-based Reinforcement Learning with a Hamiltonian Canonical ODE Network

Yao Feng,Yuhong Jiang,Hang Su,Dong Yan,Jun Zhu

Model-based reinforcement learning usually suffers from a high sample complexity in training the world model, especially for the environments with complex dynamics. To make the training for general physical environments more efficient, we introduce Hamiltonian canonical ordinary differential equations into the learning process, which inspires a novel model of neural ordinary differential auto-encoder (NODA). NODA can model the physical world by nature and is flexible to impose Hamiltonian mechanics (e.g., the dimension of the physical equations) which can further accelerate training of the environment models. It can consequentially empower an RL agent with the robust
extrapolation using a small amount of samples as well as the guarantee on the physical plausibility. Theoretically, we prove that NODA has uniform bounds for multi-step transition errors and value errors under certain conditions. Extensive experiments show that NODA can learn the environment dynamics effectively with a high sample efficiency, making it possible to facilitate reinforcement learning agents at the early stage.

****

Hinge Policy Optimization: Rethinking Policy Improvement and Reinterpreting PPO

Hsuan-Yu Yao,Ping-Chun Hsieh,Kuo-Hao Ho,Kai-Chun Hu,Liang Chun Ouyang,I-Chen Wu

Policy optimization is a fundamental principle for designing reinforcement learning algorithms, and one example is the proximal policy optimization algorithm with a clipped surrogate objective (PPO-clip), which has been popularly used in deep reinforcement learning due to its simplicity and effectiveness. Despite its superior empirical performance, PPO-clip has not been justified via theoretical proof up to date. This paper proposes to rethink policy optimization and reinterpret the theory of PPO-clip based on hinge policy optimization (HPO), called to improve policy by hinge loss in this paper. Specifically, we first identify sufficient conditions of state-wise policy improvement and then rethink policy update as solving a large-margin classification problem with hinge loss. By leveraging various types of classifiers, the proposed design opens up a whole new family of policy-based algorithms, including the PPO-clip as a special case. Based on this construct, we prove that these algorithms asymptotically attain a globally optimal policy. To our knowledge, this is the first ever that can prove global convergence to an optimal policy for a variant of PPO-clip. We corroborate the performance of a variety of HPO algorithms through experiments and an ablation study.

****

A Unified Wasserstein Distributional Robustness Framework for Adversarial Training
Anh Tuan Bui,Trung Le,Quan Hung Tran,He Zhao,Dinh Phung

It is well-known that deep neural networks (DNNs) are susceptible to adversarial attacks, exposing a severe fragility of deep learning systems. As the result, adversarial training (AT) method, by incorporating adversarial examples during training, represents a natural and effective approach to strengthen the robustness of a DNN-based classifier. However, most AT-based methods, notably PGD-AT and TRADES, typically seek a pointwise adversary that generates the worst-case adversarial example by independently perturbing each data sample, as a way to ``probe'' the vulnerability of the classifier. Arguably, there are unexplored benefits in considering such adversarial effects from an entire distribution. To this end, this paper presents a unified framework that connects Wasserstein distributional robustness with current state-of-the-art AT methods. We introduce a new Wasserstein cost function and a new series of risk functions, with which we show that standard AT methods are special cases of their counterparts in our framework. This connection leads to an intuitive relaxation and generalization of existing AT methods and facilitates the development of a new family of distributional robustness AT-based algorithms. Extensive experiments show that our distributional robustness AT algorithms robustify further their standard AT counterparts in various settings.

****

TS-BERT: A fusion model for Pre-trainning Time Series-Text Representations
Jiahao Qin,Lu Zong

There are many tasks to use news text information and stock data to predict the crisis. In the existing research, the two usually play one master and one follower in the prediction task.

Use one of the news text and the stock data as the primary information source for the prediction task and the other as the auxiliary information source.

This paper proposes a fusion model for pre-training time series-Text representations, in which news text and stock data have the same status and are treated as two different modes to describe crises. Our model has achieved the best results in the task of predicting financial crises.

****

miniF2F: a cross-system benchmark for formal Olympiad-level mathematics
Kunhao Zheng,Jesse Michael Han,Stanislas Polu

We present $\textsf{miniF2F}$, a dataset of formal Olympiad-level mathematics problems statements intended to provide a unified cross-system benchmark for neural theorem proving. The $\textsf{miniF2F}$ benchmark currently targets Metamath, Lean, Isabelle (partially) and HOL Light (partially) and consists of 488 problem statements drawn from the AIME, AMC, and the International Mathematical Olympiad (IMO), as well as material from high-school and undergraduate mathematics courses. We report baseline results using GPT-f, a neural theorem prover based on GPT-3 and provide an analysis of its performance. We intend for $\textsf{miniF2F}$ to be a community-driven effort and hope that our benchmark will help spur advances in neural theorem proving.

****

Representation mitosis in wide neural networks
Diego Doimo,Aldo Glielmo,Sebastian Goldt,Alessandro Laio

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 evidence for an underlying mechanism that we call "representation mitosis": 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. Like in a mitosis process, the number of such groups, or "clones'', increases linearly with the width of the layer, but only if the width is above a critical value. We show that a key ingredient to activate mitosis is continuing the training process until the training error is zero

****

Ensemble-in-One: Learning Ensemble within Random Gated Networks for Enhanced Adversarial Robustness

Yi Cai,Xuefei Ning,Huazhong Yang,Yu Wang

Adversarial attacks have threatened modern deep learning systems by crafting adversarial examples with small perturbations to fool the convolutional neural networks (CNNs). Ensemble training methods are promising to facilitate better adversarial robustness by diversifying the vulnerabilities among the sub-models, simultaneously maintaining comparable accuracy as standard training. Previous practices also demonstrate that enlarging the ensemble can improve the robustness. However, existing ensemble methods are with poor scalability, owing to the rapid complexity increase when including more sub-models in the ensemble. Moreover, it is usually infeasible to train or deploy an ensemble with substantial sub-models, owing to the tight hardware resource budget and latency requirement. In this work, we propose Ensemble-in-One (EIO), a simple but effective method to enlarge the ensemble within a random gated network (RGN). EIO augments the original model by replacing the parameterized layers with multi-path random gated blocks (RGBs) to construct an RGN. By diversifying the vulnerability of the numerous paths through the super-net, it provides high scalability because the paths within an RGN exponentially increase with the network depth. Our experiments demonstrate that EIO consistently outperforms previous ensemble training methods with even less computational overhead, simultaneously achieving better accuracy-robustness trade-offs than adversarial training.

****

Anomaly Transformer: Time Series Anomaly Detection with Association Discrepancy

Jiehui Xu,Haixu Wu,Jianmin Wang,Mingsheng Long

Unsupervised detection of anomaly points in time series is a challenging problem, which requires the model to derive a distinguishable criterion. Previous methods tackle the problem mainly through learning pointwise representation or pairwise association, however, neither is sufficient to reason about the intricate dynamics. Recently, Transformers have shown great power in unified modeling of pointwise representation and pairwise association, and we find that the self-attention weight distribution of each time point can embody rich association with the whole series. Our key observation is that due to the rarity of anomalies, it is extremely difficult to build nontrivial associations from abnormal points to the whole series, thereby, the anomalies' associations shall mainly concentrate on their adjacent time points. This adjacent-concentration bias implies an association-based criterion inherently distinguishable between normal and abnormal points, which we highlight through the Association Discrepancy. Technically, we propose the Anomaly Transformer with a new Anomaly-Attention mechanism to compute the association discrepancy. A minimax strategy is devised to amplify the normal-abnormal distinguishability of the association discrepancy. The Anomaly Transformer achieves state-of-the-art results on six unsupervised time series anomaly detection benchmarks of three applications: service monitoring, space & earth exploration, and water treatment.

****

Towards Model Agnostic Federated Learning Using Knowledge Distillation

Andrei Afonin,Sai Praneeth Karimireddy

Is it possible to design an universal API for federated learning using which an ad-hoc group of data-holders (agents) collaborate with each other and perform federated learning? Such an API would necessarily need to be model-agnostic i.e. make no assumption about the model architecture being used by the agents, and also cannot rely on having representative public data at hand. Knowledge distillation (KD) is the obvious tool of choice to design such protocols. However, surprisingly, we show that most natural KD-based federated learning protocols have poor performance.

To investigate this, we propose a new theoretical framework, Federated Kernel ridge regression, which can capture both model heterogeneity as well as data heterogeneity. Our analysis shows that the degradation is largely due to a fundamental limitation of knowledge distillation under data heterogeneity. We further validate our framework by analyzing and designing new protocols based on KD. Their performance on real world experiments using neural networks, though still unsatisfactory, closely matches our theoretical predictions.

****

Evolution Strategies as an Alternate Learning method for Hierarchical Reinforcement Learning

Sasha Abramowitz

This paper investigates the performance of Scalable Evolution Strategies (S-ES) as a Hierarchical Reinforcement Learning (HRL) approach. S-ES, named for its excellent scalability across many processors, was popularised by OpenAI when they showed its performance to be comparable to the state-of-the-art policy gradient methods. However, to date, S-ES has not been tested in conjunction with HRL methods, which empower temporal abstraction thus allowing agents to tackle more challenging problems. In this work, we introduce a novel method that merges S-ES and HRL, which allows S-ES to be applied to difficult problems such as simultaneous robot locomotion and navigation. We show that S-ES needed no (methodological or hyperparameter) modifications for it to be used in a hierarchical context and that its indifference to delayed rewards leads to it having competitive performance with state-of-the-art gradient-based HRL methods. This leads to a novel HRL method that achieves state-of-the-art performance, and is also comparably simple and highly scalable.

****

Deconfounding to Explanation Evaluation in Graph Neural Networks

Yingxin Wu,Xiang Wang,An Zhang,Xia Hu,Fuli Feng,Xiangnan He,Tat-Seng Chua

Explainability of graph neural networks (GNNs) aims to answer ``Why the GNN made a certain prediction?'', which is crucial to interpret the model prediction. The feature attribution framework distributes a GNN's prediction to its input features (e.g., edges), identifying an influential subgraph as the explanation. When evaluating the explanation (i.e., subgraph importance), a standard way is to audit the model prediction based on the subgraph solely. However, we argue that a distribution shift exists between the full graph and the subgraph, causing the out-of-distribution problem. Furthermore, with an in-depth causal analysis, we find the OOD effect acts as the confounder, which brings spurious associations between the subgraph importance and model prediction, making the evaluation less reliable. In this work, we propose Deconfounded Subgraph Evaluation (DSE) which assesses the causal effect of an explanatory subgraph on the model prediction. While the distribution shift is generally intractable, we employ the front-door adjustment and introduce a surrogate variable of the subgraphs. Specifically, we devise a generative model to generate the plausible surrogates that conform to the data distribution, thus approaching the unbiased estimation of subgraph importance. Empirical results demonstrate the effectiveness of DSE in terms of explanation fidelity.

****

Acceleration of Federated Learning with Alleviated Forgetting in Local Training

Chencheng Xu,Zhiwei Hong,Minlie Huang,Tao Jiang

Federated learning (FL) enables distributed optimization of machine learning models while protecting privacy by independently training local models on each client and then aggregating parameters on a central server, thereby producing an effective global model. Although a variety of FL algorithms have been proposed, their training efficiency remains low when the data are not independently and identically distributed (non-i.i.d.) across different clients. We observe that the slow convergence rates of the existing methods are (at least partially) caused by the catastrophic forgetting issue during the local training stage on each individual client, which leads to a large increase in the loss function concerning the previous training data provided at other clients. Here, we propose FedReg, an a lgorithm to accelerate FL with alleviated knowledge forgetting in the local training stage by regularizing locally trained parameters with the loss on generated pseudo data, which encode the knowledge of previous training data learned by the global model. Our comprehensive experiments demonstrate that FedReg not only significantly improves the convergence rate of FL, especially when the neural network architecture is deep and the clients' data are extremely non-i.i.d., but is also able to protect privacy better in classification problems and more robust against gradient inversion attacks.

****

Discovering Invariant Rationales for Graph Neural Networks

Yingxin Wu,Xiang Wang,An Zhang,Xiangnan He,Tat-Seng Chua

Intrinsic interpretability of graph neural networks (GNNs) is to find a small subset of the input graph's features --- rationale --- which guides the model prediction. Unfortunately, the leading rationalization models often rely on data biases, especially shortcut features, to compose rationales and make predictions without probing the critical and causal patterns. Moreover, such data biases easily change outside the training distribution. As a result, these models suffer from a huge drop in interpretability and predictive performance on out-of-distribution data. In this work, we propose a new strategy of discovering invariant rationale (DIR) to construct intrinsically interpretable GNNs. It conducts interventions on the training distribution to create multiple interventional distributions. Then it approaches the causal rationales that are invariant across different distributions while filtering out the spurious patterns that are unstable. Experiments on both synthetic and real-world datasets validate the superiority of our DIR in terms of interpretability and generalization ability on graph classification over the leading baselines. Code and datasets are available at https://github.com/Wuyxin/DIR-GNN.

****

EAT-C: Environment-Adversarial sub-Task Curriculum for Efficient Reinforcement Learning

Shuang Ao,Tianyi Zhou,Jing Jiang,Guodong Long,Xuan Song,Chengqi Zhang

Reinforcement learning (RL)'s efficiency can drastically degrade on long-horizon tasks due to sparse rewards and the RL policy can be fragile to small changes in deployed environments. To improve RL's efficiency and generalization to varying environments, we study how to automatically generate a curriculum of tasks with coupled environments for RL. To this end, we train two curriculum policies together with RL: (1) a co-operative planning policy recursively decomposing a hard task into coarse-to-fine sub-task sequences as a tree; and (2) an adversarial policy modifying the environment (e.g., position/size of obstacles) in each sub-task. They are complementary in acquiring more informative feedback for RL: the planning policy provides dense reward of finishing easier sub-tasks while the environment policy modifies these sub-tasks to be adequately challenging and diverse so the RL agent can quickly adapt to different tasks/environments. On the other hand, they are trained using the RL agent's dense feedback on sub-tasks so the sub-task curriculum keeps adaptive to the agent's progress via this ``iterative mutual-boosting'' scheme. Moreover, the sub-task tree naturally enables an easy-to-hard curriculum for every policy: its top-down construction gradually increases sub-tasks the planning policy needs to generate, while the adversarial training between the environment policy and the RL policy follows a bottom-up traversal that starts from a dense sequence of easier sub-tasks allowing more frequent modifications to the environment. Therefore, jointly training the three policies leads to efficient RL guided by a curriculum progressively improving the sparse reward and generalization. We compare our method with popular RL/planning approaches targeting similar problems and the ones with environment generators or adversarial agents. Thorough experiments on diverse benchmark tasks demonstrate significant advantages of our method on improving RL's efficiency and generalization.

****

ZenDet: Revisiting Efficient Object Detection Backbones from Zero-Shot Neural Architecture Search

Zhenhong Sun,Ming Lin,Zhiyu Tan,Xiuyu Sun,Rong Jin

In object detection models, the detection backbone consumes more than half of the overall inference cost. Recent researches attempt to reduce this cost by optimizing the backbone architecture with the help of Neural Architecture Search (NAS). However, existing NAS methods for object detection require hundreds to thousands of GPU hours of searching, making them impractical in fast-paced research and development. In this work, we propose a novel zero-shot NAS method to address this issue. The proposed method, named ZenDet, automatically designs efficient detection backbones without training network parameters, reducing the architecture design cost to nearly zero yet delivering the state-of-the-art (SOTA) performance. Under the hood, ZenDet maximizes the differential entropy of detection backbones, leading to a better feature extractor for object detection under the same computational budgets. After merely one GPU day of fully automatic design, ZenDet innovates SOTA detection backbones on multiple detection benchmark datasets with little human intervention. Comparing to ResNet-50 backbone, ZenDet is $+2.0\%$ better in mAP when using the same amount of FLOPs/parameters and is 1.54 times faster on NVIDIA V100 at the same mAP. Code and pre-trained models will be released after publication.

****

Closed-Loop Data Transcription to an LDR via Minimaxing Rate Reduction

Xili Dai,Shengbang Tong,Mingyang Li,Ziyang Wu,Kwan Ho Ryan Chan,Pengyuan Zhai,Yaodong Yu,Michael Psenka,Xiaojun Yuan,Heung-Yeung Shum,Yi Ma

This work proposes a new computational framework for automatically learning a closed-loop transcription between multi-class multi-dimensional data and a linear discriminative representation (LDR) that consists of multiple multi-dimensional linear subspaces. In particular, we argue that the optimal encoding and decoding mappings sought can be formulated as the equilibrium point of a two-player minimax game between the encoder and decoder. A natural utility function for this game is the so-called rate reduction, a simple information-theoretic measure for distances between mixtures of subspace-like Gaussians in the feature space. Our formulation avoids expensive evaluating and minimizing approximated distances between arbitrary distributions in either the data space or the feature space. To a large extent, conceptually and computationally this new formulation unifies the benefits of Auto-Encoding and GAN and naturally extends them to the settings of learning a both discriminative and generative representation for complex multi-class and multi-dimensional real-world data. Our extensive experiments on many benchmark datasets demonstrate tremendous potential of this framework: under fair comparison, visual quality of the learned decoder and classification performance of the encoder is competitive and often better than existing methods based on GAN, VAE or a combination of both.

****

Representing Mixtures of Word Embeddings with Mixtures of Topic Embeddings

dongsheng wang,Dan dan Guo,He Zhao,Huangjie Zheng,Korawat Tanwisuth,Bo Chen,Mingyuan Zhou

A topic model is often formulated as a generative model that explains how each word of a document is generated given a set of topics and document-specific topic proportions. It is focused on capturing the word co-occurrences in a document and hence often suffers from poor performance in analyzing short documents. In addition, its parameter estimation often relies on approximate posterior inference that is either not scalable or suffering from large approximation error. This paper introduces a new topic-modeling framework where each document is viewed as a set of word embedding vectors and each topic is modeled as an embedding vector in the same embedding space. Embedding the words and topics in the same vector space, we define a method to measure the semantic difference between the embedding vectors of the words of a document and these of the topics, and optimize the topic embeddings to minimize the expected difference over all documents. Experiments on text analysis demonstrate that the proposed method, which is amenable to mini-batch stochastic gradient descent based optimization and hence scalable to big corpora, provides competitive performance in discovering more coherent and diverse topics and extracting better document representations.

****
Generative Modeling with Optimal Transport Maps
Litu Rout,Alexander Korotin,Evgeny Burnaev

With the discovery of Wasserstein GANs, Optimal Transport (OT) has become a powerful tool for large-scale generative modeling tasks. In these tasks, OT cost is typically used as the loss for training GANs. In contrast to this approach, we show that the OT map itself can be used as a generative model, providing comparable performance. Previous analogous approaches consider OT maps as generative models only in the latent spaces due to their poor performance in the original high-dimensional ambient space. In contrast, we apply OT maps directly in the ambient space, e.g., a space of high-dimensional images. First, we derive a min-max optimization algorithm to efficiently compute OT maps for the quadratic cost (Wasserstein-2 distance). Next, we extend the approach to the case when the input and output distributions are located in the spaces of different dimensions and derive error bounds for the computed OT map. We evaluate the algorithm on image generation and unpaired image restoration tasks. In particular, we consider denoising, colorization, and inpainting, where the optimality of the restoration map is a desired attribute, since the output (restored) image is expected to be close to the input (degraded) one.

****
Focus on the Common Good: Group Distributional Robustness Follows
Vihari Piratla,Praneeth Netrapalli,Sunita Sarawagi

We consider the problem of training a classification model with group annotated training data. Recent work has established that, if there is distribution shift across different groups, models trained using the standard empirical risk minimization (ERM) objective suffer from poor performance on minority groups and that group distributionally robust optimization (Group-DRO) objective is a better alternative. The starting point of this paper is the observation that though Group-DRO performs better than ERM on minority groups for some benchmark datasets, there are several other datasets where it performs much worse than ERM. Inspired by ideas from the closely related problem of domain generalization, this paper proposes a new and simple algorithm that explicitly encourages learning of features that are shared across various groups. The key insight behind our proposed algorithm is that while Group-DRO focuses on groups with worst regularized loss, focusing instead, on groups that enable better performance even on other groups, could lead to learning of shared/common features, thereby enhancing minority performance beyond what is achieved by Group-DRO. Empirically, we show that our proposed algorithm matches or achieves better performance compared to strong contemporary baselines including ERM and Group-DRO on standard benchmarks on both minority groups and across all groups. Theoretically, we show that the proposed algorithm is a descent method and finds first order stationary points of smooth nonconvex functions.

****
Greedy Bayesian Posterior Approximation with Deep Ensembles
Aleksei Tiulpin,Matthew B. Blaschko

Ensembles of independently trained neural networks are a state-of-the-art approach to estimate predictive uncertainty in Deep Learning, and can be interpreted as an approximation of the posterior distribution via a mixture of delta functions. The training of ensembles relies on non-convexity of the loss landscape and random initialization of their individual members, making the resulting posterior approximation uncontrolled. This paper proposes a novel and principled method to tackle this limitation, minimizing an f-divergence between the true posterior and a kernel density estimator in a function space. We analyze this objective from a combinatorial point of view, and show that it is submodular with respect to mixture components for any f. Subsequently, we consider the problem of greedy ensemble construction, and from the marginal gain of the total objective, we derive a novel diversity term for ensemble methods. The performance of our approach is demonstrated on computer vision out-of-distribution detection benchmarks in a range of architectures trained on multiple datasets. The source code of our method is made publicly available.

****

A Topological View of Rule Learning in Knowledge Graphs

Zuoyu Yan,Tengfei Ma,Liangcai Gao,Zhi Tang,Chao Chen

Inductive relation prediction is an important learning task for knowledge graph completion. One can use the existence of rules, namely a sequence of relations, to predict the relation between two entities. Previous works view rules as paths and primarily focus on the searching of paths between entities. The space of paths is huge, and one has to sacrifice either efficiency or accuracy. In this paper, we consider rules in knowledge graphs as cycles and show that the space of cycles has a unique structure based on the theory of algebraic topology. By exploring the linear structure of the cycle space, we can improve the searching efficiency of rules. We propose to collect cycle bases that span the space of cycles. We build a novel GNN framework on the collected cycles to learn the representations of cycles, and to predict the existence/non-existence of a relation. Our method achieves state-of-the-art performance on benchmarks.

****

Omni-Scale CNNs: a simple and effective kernel size configuration for time series classification

Wensi Tang,Guodong Long,Lu Liu,Tianyi Zhou,Michael Blumenstein,Jing Jiang

The size of the receptive field has been one of the most important factors for One Dimensional Convolutional Neural Networks (1D-CNNs) on time series classification tasks. Large efforts have been taken to choose the appropriate receptive field size, for it has a huge influence on the performance and differs significantly for each dataset. In this paper, we propose an Omni-Scale block (OS-block) for 1D-CNNs, where the kernel sizes are set by a simple and universal rule. OS-block can efficiently cover the best size of the receptive field across different datasets. This set of kernel sizes consists of multiple prime numbers according to the length of the time series. We experimentally show 1D-CNNs built from OS-block can consistently achieve the state-of-the-art accuracy with a smaller model size on five time series benchmarks, including both univariate and multivariate data from multiple domains. Comprehensive analysis and ablation studies shed light on how our rule finds the best receptive field size and demonstrate the consistency of our OS-block for multiple 1D-CNN structures.

****

SPARK: co-exploring model SPArsity and low-RanKness for compact neural networks

Wanzhao Yang,Miao Yin,Yang Sui,Bo Yuan

Sparsification and low-rank decomposition are two important techniques for deep neural network (DNN) compression. To date, these two popular yet distinct approaches are typically used in a separate way; while their efficient integration for better compression performance is little explored. In this paper we perform systematic co-exploration on the model sparsity and low-rankness towards compact neural networks. We first investigate and analyze several important design factors for the joint pruning and low-rank factorization, including operational sequence, low-rank format, and optimization objective. Based on the observations and outcomes from our analysis, we then propose SPARK, a unified DNN compression framework that can simultaneously capture model SPArsity and low-RanKness in an efficient way. Empirical experiments demonstrate very promising performance of our proposed solution. Notably, on CIFAR-10 dataset, our approach can bring 1.25%, 1.02% and 0.16% accuracy increase over the baseline ResNet-20, ResNet-56 and DenseNet-40 models, respectively, and meanwhile the storage and computational costs are reduced by 70.4% and 71.1% (for ResNet-20), 37.5% and 39.3% (for ResNet-56) and 52.4% and 61.3% (for DenseNet-40), respectively. On ImageNet dataset, our approach can enable 0.52% accuracy increase over baseline model with 48.7% fewer parameters.

****

Ada-NETS: Face Clustering via Adaptive Neighbour Discovery in the Structure Space

Yaohua Wang,Yaobin Zhang,Fangyi Zhang,Senzhang Wang,Ming Lin,YuQi Zhang,Xiuyu Sun

Face clustering has attracted rising research interest recently to take advantag e of massive amounts of face images on the web. State-of-the-art performance has been achieved by Graph Convolutional Networks (GCN) due to their powerful representation capacity. However, existing GCN-based methods build face graphs mainly according to $k$NN relations in the feature space, which may lead to a lot of noise edges connecting two faces of different classes. The face features will be polluted when messages pass along these noise edges, thus degrading the performance of GCNs. In this paper, a novel algorithm named Ada-NETS is proposed to cluster faces by constructing clean graphs for GCNs. In Ada-NETS, each face is transformed to a new structure space, obtaining robust features by considering face features of the neighbour images. Then, an adaptive neighbour discovery strategy is proposed to determine a proper number of edges connecting to each face image. It significantly reduces the noise edges while maintaining the good ones to build a graph with clean yet rich edges for GCNs to cluster faces. Experiments on multiple public clustering datasets show that Ada-NETS significantly outperforms current state-of-the-art methods, proving its superiority and generalization. Code is available at https://github.com/damo-cv/Ada-NETS.

****

Decoupled Adaptation for Cross-Domain Object Detection
Junguang Jiang,Baixu Chen,Jianmin Wang,Mingsheng Long

Cross-domain object detection is more challenging than object classification since multiple objects exist in an image and the location of each object is unknown in the unlabeled target domain. As a result, when we adapt features of different objects to enhance the transferability of the detector, the features of the foreground and the background are easy to be confused, which may hurt the discriminability of the detector. Besides, previous methods focused on category adaptation but ignored another important part for object detection, i.e., the adaptation on bounding box regression. To this end, we propose D-adapt, namely Decoupled Adaptation, to decouple the adversarial adaptation and the training of the detector. Besides, we fill the blank of regression domain adaptation in object detection by introducing a bounding box adaptor. Experiments show that \textit{D-adapt} achieves state-of-the-art results on four cross-domain object detection tasks and yields 17\% and 21\% relative improvement on benchmark datasets Clipart1k and Comic2k in particular.

****

Rethinking Network Design and Local Geometry in Point Cloud: A Simple Residual MLP Framework
Xu Ma,Can Qin,Haoxuan You,Haoxi Ran,Yun Fu

Point cloud analysis is challenging due to irregularity and unordered data structure. To capture the 3D geometries, prior works mainly rely on exploring sophisticated local geometric extractors, using convolution, graph, or attention mechanisms. These methods, however, incur unfavorable latency during inference and the performance saturates over the past few years. In this paper, we present an ovel perspective on this task. We find detailed local geometrical informationprobably is not the key to point cloud analysis – we introduce a pure residual MLP network, called PointMLP, which integrates no local geometrical extractors but still performs very competitively. Equipped with a proposed lightweight geometric-affine module to stabilize the training, PointMLP delivers the new state-of-the-art on multiple datasets. On the real-world ScanObjectNN dataset, our method even surpasses the prior best method by 3.3% accuracy. We emphasize PointMLP achieves this strong performance without any sophisticated operations, hence leading to a prominent inference speed. Compared to most recent CurveNet, PointMLP trains 2× faster, tests 7× faster, and is more accurate on ModelNet40 benchmark. We hope our PointMLP may help the community towards a better understanding of point cloud analysis. The code is available at https://github.com/ma-xu/pointMLP-pytorch.

****

Progressive Distillation for Fast Sampling of Diffusion Models
Tim Salimans,Jonathan Ho

Diffusion models have recently shown great promise for generative modeling, outperforming GANs on perceptual quality and autoregressive models at density estimation. A remaining downside is their slow sampling time: generating high quality samples takes many hundreds or thousands of model evaluations. Here we make two contributions to help eliminate this downside: First, we present new parameterizations of diffusion models that provide increased stability when using few sampling steps, compared to models in the literature. Second, we present a method to distill a trained deterministic diffusion sampler, using many steps, into a new diffusion model that takes half as many sampling steps. We then keep progressively applying this distillation procedure to our model, halving the number of required sampling steps each time. On standard image generation benchmarks like CIFAR-10, ImageNet, and LSUN, we start out with (near) state-of-the-art samplers taking 1024 or 8192 steps, and are able to distill down to models taking as little as 4 steps without losing much perceptual quality; achieving, for example, a FID of 3.0 on CIFAR-10 in 4 steps. Finally, we show that the full progressive distillation procedure does not take more time than it takes to train the original model, thus representing an efficient solution for generative modeling using diffusion at both train and test time.

****

Unconditional Diffusion Guidance

Jonathan Ho,Tim Salimans

Classifier guidance is a recently introduced method to trade off mode coverage and sample fidelity in conditional diffusion models post training, in the same spirit as low temperature sampling or truncation in other types of generative models. Classifier guidance combines the score estimate of a diffusion model with the gradient of an image classifier and thereby requires training an image classifier separate from the diffusion model. It also raises the question of whether guidance can be performed without a classifier. We show that guidance can be indeed performed by a pure generative model without such a classifier: in what we call unconditional guidance, we jointly train a conditional and an unconditional diffusion model, and we combine the resulting conditional and unconditional score estimates to attain a trade-off between sample quality and diversity similar to that obtained using classifier guidance.

****

New Insights on Reducing Abrupt Representation Change in Online Continual Learning

Lucas Caccia,Rahaf Aljundi,Nader Asadi,Tinne Tuytelaars,Joelle Pineau,Eugene Belilovsky

In the online continual learning paradigm, agents must learn from a changing distribution while respecting memory and compute constraints. Experience Replay (ER), where a small subset of past data is stored and replayed alongside new data, has emerged as a simple and effective learning strategy. In this work, we focus on the change in representations of observed data that arises when previously unobserved classes appear in the incoming data stream, and new classes must be distinguished from previous ones. We shed new light on this question by showing that applying ER causes the newly added classes’ representations to overlap significantly with the previous classes, leading to highly disruptive parameter updates. Based on this empirical analysis, we propose a new method which mitigates this issue by shielding the learned representations from drastic adaptation to accommodate new classes. We show that using an asymmetric update rule pushes new classes to adapt to the older ones (rather than the reverse), which is more effective especially at task boundaries, where much of the forgetting typically occurs. Empirical results show significant gains over strong baselines on standard continual learning benchmarks.

****

On Anytime Learning at Macroscale

Lucas Caccia,Jing Xu,Myle Ott,MarcAurelio Ranzato,Ludovic Denoyer

Classical machine learning frameworks assume access to a possibly large dataset in order to train a predictive model. In many practical applications however, data does not arrive all at once, but in batches over time. This creates a natural trade-off between accuracy of a model and time to obtain such a model. A greedy predictor could produce non-trivial predictions by immediately training on batches as soon as these become available but, it may also make sub-optimal use of future data. On the other hand, a tardy predictor could wait for a long time to aggregate several batches into a larger dataset, but ultimately deliver a much better performance. In this work, we consider such a streaming learning setting, which we dub {\em anytime learning at macroscale} (ALMA). It is an instance of anytime learning applied not at the level of a single chunk of data, but at the level of the entire sequence of large batches. We first formalize this learning setting, we then introduce metrics to assess how well learners perform on the given task for a given memory and compute budget, and finally we test about thirty baseline approaches on three standard benchmarks repurposed for anytime learning at macroscale. Our findings indicate that no model strikes the best trade-off across the board. While replay-based methods attain the lowest error rate, they also incur in a 5 to 10 times increase of compute. Approaches that grow capacity over time do offer better scaling in terms of training flops, but they also underperform simpler ensembling methods in terms of error rate. Overall, ALMA offers both a good abstraction of the typical learning setting faced everyday by practitioners, and a set of unsolved modeling problems for those interested in efficient learning of dynamic models.

****

Demystifying Batch Normalization in ReLU Networks: Equivalent Convex Optimization Models and Implicit Regularization

Tolga Ergen,Arda Sahiner,Batu Ozturkler,John M. Pauly,Morteza Mardani,Mert Pilanci

Batch Normalization (BN) is a commonly used technique to accelerate and stabilize training of deep neural networks. Despite its empirical success, a full theoretical understanding of BN is yet to be developed. In this work, we analyze BN through the lens of convex optimization. We introduce an analytic framework based on convex duality to obtain exact convex representations of weight-decay regularized ReLU networks with BN, which can be trained in polynomial-time. Our analyses also show that optimal layer weights can be obtained as simple closed-form formulas in the high-dimensional and/or overparameterized regimes. Furthermore, we find that Gradient Descent provides an algorithmic bias effect on the standard non-convex BN network, and we design an approach to explicitly encode this implicit regularization into the convex objective. Experiments with CIFAR image classification highlight the effectiveness of this explicit regularization for mimicking and substantially improving the performance of standard BN networks.

****

Associated Learning: an Alternative to End-to-End Backpropagation that Works on CNN, RNN, and Transformer

Dennis Y.H. Wu,Dinan Lin,Vincent Chen,Hung-Hsuan Chen

This paper studies Associate Learning (AL), an alternative methodology to the end-to-end backpropagation (BP). We introduce the workflow to convert a neural network into a proper structure such that AL can be used to learn the weights for various types of neural networks. We compared AL and BP on some of the most successful types of neural networks -- Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), and Transformer. Experimental results show that AL consistently outperforms BP on various open datasets. We discuss possible reasons for AL's success and its limitations.

****

MetaShift: A Dataset of Datasets for Evaluating Contextual Distribution Shifts and Training Conflicts

Weixin Liang,James Zou

Understanding the performance of machine learning models across diverse data distributions is critically important for reliable applications. Motivated by this, there is a growing focus on curating benchmark datasets that capture distribution shifts. While valuable, the existing benchmarks are limited in that many of them only contain a small number of shifts and they lack systematic annotation about what is different across different shifts. We present MetaShift—a collection of 12,868 sets of natural images across 410 classes—to address this challenge. We leverage the natural heterogeneity of Visual Genome and its annotations to construct MetaShift. The key construction idea is to cluster images using its meta data, which provides context for each image (e.g. “cats with cars” or “cats in bathroom”) that represent distinct data distributions. MetaShift has two important benefits: first, it contains orders of magnitude more natural data shifts than previously available. Second, it provides explicit explanations of what is unique about each of its data sets and a distance score that measures the amount of distribution shift between any two of its data sets. We demonstrate the utility of MetaShift in benchmarking several recent proposals for training models to be robust to data shifts. We find that the simple empirical risk minimization performs the best when shifts are moderate and no method had a systematic advantage for large shifts. We also show how MetaShift can help to visualize conflicts between data subsets during model training.

****

Text-Driven Image Manipulation via Semantic-Aware Knowledge Transfer

Ziqi Zhang,Cheng Deng,Kun Wei,Xu Yang

Semantic-level facial attribute transfer is a special task to edit facial attribute, when reference images are viewed as conditions to control the image editing. In order to achieve better performance, semantic-level facial attribute transfer needs to fulfil two requirements: (1) specific attributes extracted from reference face should be precisely transferred to target face; (2) irrelevant information should be completely retained after transferring. Some existing methods locate and modify local support regions of facial images, which are not effective when editing global attributes; the other methods disentangle the latent code as different attribute-relevant parts, which may transfer redundant knowledge to target faces. In this paper, we first propose a novel text-driven directional latent mapping network with semantic direction consistency (SDC) constrain to explore the latent semantic space for effective attribute editing, leveraging the semantic-aware knowledge of Contrastive Language-Image Pre-training (CLIP) model as guidance. This latent space manipulation strategy is designed to disentangle the facial attribute, removing the redundant knowledge in the transfer process. And on this basis, a novel attribute transfer method, named semantic directional decomposition network (SDD-Net), is proposed to achieve semantic-level facial attribute transfer by latent semantic direction decomposition, improving the interpretability and editability of our method. Extensive experiments on CelebA-HQ dataset show that our method achieves impressive performance over the state-of-the-art methods.

****

Cycle monotonicity of adversarial attacks for optimal domain adaptation

Arip Asadulaev,Vitaly Shutov,Alexander Korotin,Alexander Panfilov,Andrey Filchenkov

We reveal an intriguing connection between adversarial attacks and cycle monotone maps, also known as optimal transport maps. Based on this finding, we developed a novel method named source fiction for semi-supervised optimal transport-based domain adaptation. In our algorithm, instead of mapping from target to the source domain, optimal transport maps target samples to the set of adversarial examples. The trick is that these adversarial examples are labeled target samples perturbed to look like source samples for the source domain classifier. Due to the cycle monotonicity of adversarial attacks, optimal transport can naturally approximate this transformation. We conduct experiments on various datasets and show that our method can notably improve the performance of optimal transport methods in semi-supervised domain adaptation.

****

Pairwise Adversarial Training for Unsupervised Class-imbalanced Domain Adaptation

Weili Shi,Ronghang Zhu,Sheng Li

Unsupervised domain adaptation (UDA) has become an appealing approach for knowledge transfer from a labeled source domain to an unlabeled target domain. However, when the classes in source and target domains are imbalanced, most existing UDA methods experience significant performance drop, as the decision boundary usually favors the majority classes. Some recent class-imbalanced domain adaptation (CDA) methods aim to tackle the challenge of biased label distribution by exploi ting pseudo-labeled target data during training process. However, these methods may be challenged with the problem of unreliable pseudo labels and error accumulation during training. In this paper, we propose a pairwise adversarial training approach to augment training data for unsupervised class-imbalanced domain adaptation. Unlike conventional adversarial training in which the adversarial samples are obtained from the ℓ_p ball of the original data, we obtain the semantic adversarial samples from the interpolated line of the aligned pair-wise samples from source domain and target domain. Experimental results and ablation study show that our method can achieve considerable improvements on the CDA benchmarks compared with the state-of-art methods focusing on the same problem.

****

Cognitively Inspired Learning of Incremental Drifting Concepts

Mohammad Rostami,Aram Galstyan

Humans continually expand their learned knowledge to new domains and learn new concepts without any interference with past learned experiences. In contrast, machine learning models perform poorly in a continual learning setting, where input data distribution changes over time. Inspired by the nervous system learning mechanisms, we develop a computational model that enables a deep neural network to learn new concepts and expand its learned knowledge to new domains incrementally in a continual learning setting. We rely on the Parallel Distributed Processing theory to encode abstract concepts in an embedding space in terms of a multimodal distribution. This embedding space is modeled by internal data representations in a hidden network layer. We also leverage the Complementary Learning Systems theory to equip the model with a memory mechanism to overcome catastrophic forgetting through implementing pseudo-rehearsal. Our model can generate pseudo-data points for experience replay and accumulate new experiences to past learned experiences without causing cross-task interference.

****

Automated Mobile Attention KPConv Networks via A Wide & Deep Predictor

Tunhou Zhang,Mingyuan Ma,Feng Yan,Hai Li,Yiran Chen

Kernel Point Convolution (KPConv) achieves cutting-edge performance on 3D point cloud applications. Unfortunately, the large size of KPConv network limits its usage in mobile scenarios. In addition, we observe that KPConv ignores the kernel relationship and treats each kernel point equally when formulating neighbor-kernel correlation via Euclidean distance. This leads to a weak representation power. To mitigate the above issues, we propose a module named Mobile Attention Kernel Point Convolution (MAKPConv) to improve the efficiency and quality of KPConv. MAKPConv employs a depthwise kernel to reduce resource consumption and re-calibrates the contribution of kernel points towards each neighbor point via Neighbor-Kernel attention to improve representation power. Furthermore, we capitalize Inverted Residual Bottleneck (IRB) to craft a design space and employ a predictor-based Neural Architecture Search (NAS) approach to automate the design of efficient 3D networks based on MAKPConv. To fully exploit the immense design space via an accurate predictor, we identify the importance of carrying feature engineering on searchable features to improve neural architecture representations and propose a Wide & Deep Predictor to unify dense and sparse neural architecture representations for lower error in performance prediction. Experimental evaluations show that our NAS-crafted MAKPConv network uses 96% fewer parameters on 3D point cloud classification and segmentation benchmarks with better performance. Compared with state-of-the-art NAS-crafted model SPVNAS, our NAS-crafted MAKPConv network achieves ~1% better mIOU with 83% fewer parameters and 52% fewer Multiply-Accumulates.

****

Evaluating Deep Graph Neural Networks

Wentao Zhang,Zeang Sheng,Jiang Yuezihan,Yikuan Xia,Jun Gao,Zhi Yang,Bin CUI

Graph Neural Networks (GNNs) have already been widely applied in various graph mining tasks. However, most GNNs only have shallow architectures, which limits performance improvement. In this paper, we conduct a systematic experimental evaluation on the fundamental limitations of current architecture designs. Based on t he experimental results, we answer the following two essential questions: (1) what actually leads to the compromised performance of deep GNNs; (2) how to build deep GNNs. The answers to the above questions provide empirical insights and guidelines for researchers to design deep GNNs. Further, we present Deep Graph Multi-Layer Perceptron (DGMLP), a powerful approach implementing our proposed guidelines. Experimental results demonstrate three advantages of DGMLP: 1) high accuracy -- it achieves state-of-the-art node classification performance on various datasets; 2) high flexibility -- it can flexibly choose different propagation and transformation depths according to certain graph properties; 3) high scalability and efficiency -- it supports fast training on large-scale graphs.

****

FP-DETR: Detection Transformer Advanced by Fully Pre-training

Wen Wang,Yang Cao,Jing Zhang,Dacheng Tao

Large-scale pre-training has proven to be effective for visual representation learning on downstream tasks, especially for improving robustness and generalization. However, the recently developed detection transformers only employ pre-training on its backbone while leaving the key component, i.e., a 12-layer transformer, being trained from scratch, which prevents the model from above benefits. This separated training paradigm is mainly caused by the discrepancy between the upstream and downstream tasks. To mitigate the issue, we propose FP-DETR, a new method that Fully Pre-Trains an encoder-only transformer and smoothly fine-tunes it for object detection via a task adapter. Inspired by the success of textual prompts in NLP, we treat query positional embeddings as visual prompts to help the model attend to the target area (prompting) and recognize the object. To this end, we propose the task adapter which leverages self-attention to model the contextual relation between object query embedding. Experiments on the challenging COCO dataset demonstrate that our FP-DETR achieves competitive performance. Moreover, it enjoys better robustness to common corruptions and generalization to small-size datasets than state-of-the-art detection transformers. Code will be made publicly available at $\url{https://github.com/encounter1997/FP-DETR}$.

****

Towards Scheduling Federated Deep Learning using Meta-Gradients for Inter-Hospital Learning

Rasheed El-Bouri,Tingting Zhu,David A. Clifton

Given the abundance and ease of access of personal data today, individual privacy has become of paramount importance, particularly in the healthcare domain. In this work, we aim to utilise patient data extracted from multiple hospital data centres to train a machine learning model without sacrificing patient privacy. We develop a scheduling algorithm in conjunction with a student-teacher algorithm that is deployed in a federated manner. This allows a central model to learn from batches of data at each federal node. The teacher acts between data centres to update the main task (student) algorithm using the data that is stored in the various data centres. We show that the scheduler, trained using meta-gradients, can effectively organise training and as a result train a machine learning model on a diverse dataset without needing explicit access to the patient data. We achieve state-of-the-art performance and show how our method overcomes some of the problems faced in the federated learning such as node poisoning. We further show how the scheduler can be used as a mechanism for transfer learning, allowing different teachers to work together in training a student for state-of-the-art performance.

****

Efficient and Differentiable Conformal Prediction with General Function Classes

Yu Bai,Song Mei,Huan Wang,Yingbo Zhou,Caiming Xiong

Quantifying the data uncertainty in learning tasks is often done by learning a prediction interval or prediction set of the label given the input. Two commonly desired properties for learned prediction sets are \emph{valid coverage} and \emph{good efficiency} (such as low length or low cardinality). Conformal prediction is a powerful technique for learning prediction sets with valid coverage, yet by default its conformalization step only learns a single parameter, and does not optimize the efficiency over more expressive function classes.

In this paper, we propose a generalization of conformal prediction to multiple learnable parameters, by considering the constrained empirical risk minimization (ERM) problem of finding the most efficient prediction set subject to valid empirical coverage. This meta-algorithm generalizes existing conformal prediction algorithms, and we show that it achieves approximate valid population coverage and near-optimal efficiency within class, whenever the function class in the conformalization step is low-capacity in a certain sense. Next, this ERM problem is challenging to optimize as it involves a non-differentiable coverage constraint. We develop a gradient-based algorithm for it by approximating the original constrained ERM using differentiable surrogate losses and Lagrangians. Experiments show that our algorithm is able to learn valid prediction sets and improve the efficiency significantly over existing approaches in several applications such as prediction intervals with improved length, minimum-volume prediction sets for multi-output regression, and label prediction sets for image classification.
****

On the One-sided Convergence of Adam-type Algorithms in Non-convex Non-concave Min-max Optimization
Zehao Dou,Yuanzhi Li
Adam-type methods, the extension of adaptive gradient methods, have shown great performance in the training of both supervised and unsupervised machine learning models. In particular, Adam-type optimizers have been widely used empirically as the default tool for training generative adversarial networks (GANs). On the theory side, however, despite the existence of theoretical results showing the efficiency of Adam-type methods in minimization problems, the reason of their wonderful performance still remains absent in GAN's training. In existing works, the fast convergence has long been considered as one of the most important reasons and multiple works have been proposed to give a theoretical guarantee of the convergence to a critical point of min-max optimization algorithms under certain assumptions. In this paper, we firstly argue empirically that in GAN's training, Adam does not converge to a critical point even upon successful training: Only the generator is converging while the discriminator's gradient norm remains high throughout the training. We name this one-sided convergence. Then we bridge the gap between experiments and theory by showing that Adam-type algorithms provably converge to a one-sided first order stationary points in min-max optimization problems under the one-sided MVI condition. We also empirically verify that such one-sided MVI condition is satisfied for standard GANs after trained over standard data sets. To the best of our knowledge, this is the very first result which provides an empirical observation and a strict theoretical guarantee on the one-sided convergence of Adam-type algorithms in min-max optimization.
****

Translatotron 2: Robust direct speech-to-speech translation
Ye Jia,Michelle Tadmor Ramanovich,Tal Remez,Roi Pomerantz
We present Translatotron 2, a neural direct speech-to-speech translation model that can be trained end-to-end. Translatotron 2 consists of a speech encoder, a phoneme decoder, a mel-spectrogram synthesizer, and an attention module that connects all the previous three components. Experimental results suggest that Translatotron 2 outperforms the original Translatotron by a large margin in terms of translation quality and predicted speech naturalness, and drastically improves the robustness of the predicted speech by mitigating over-generation, such as babbling or long pause. We also propose a new method for retaining the source speaker's voice in the translated speech. The trained model is restricted to retain the source speaker's voice, but unlike the original Translatotron, it is not able to generate speech in a different speaker's voice, making the model more robust for production deployment, by mitigating potential misuse for creating spoofing audio artifacts. When the new method is used together with a simple concatenation-based data augmentation, the trained Translatotron 2 model is able to retain each speaker's voice for input with speaker turns.
****

Safe Neurosymbolic Learning with Differentiable Symbolic Execution
Chenxi Yang,Swarat Chaudhuri

We study the problem of learning verifiably safe parameters for programs that use neural networks as well as symbolic, human-written code. Such neurosymbolic programs arise in many safety-critical domains. However, because they need not be differentiable, it is hard to learn their parameters using existing gradient-based approaches to safe learning. Our method, Differentiable Symbolic Execution (DSE), samples control flow paths in a program, symbolically constructs worst-case "safety loss" along these paths, and backpropagates the gradients of these losses through program operations using a generalization of the REINFORCE estimator. We evaluate the method on a mix of synthetic tasks and real-world benchmarks. Our experiments show that DSE significantly outperforms the state-of-the-art DiffAI method on these tasks.
****
Provable Hierarchy-Based Meta-Reinforcement Learning
Kurtland Chua,Qi Lei,Jason D. Lee
Hierarchical reinforcement learning (HRL) has seen widespread interest as an approach to tractable learning of complex modular behaviors. However, existing work either assume access to expert-constructed hierarchies, or use hierarchy-learning heuristics with no provable guarantees. To address this gap, we analyze HRL in the meta-RL setting, where a learner learns latent hierarchical structure during meta-training for use in a downstream task. We consider a tabular setting where natural hierarchical structure is embedded in the transition dynamics. Analogous to supervised meta-learning theory, we provide “diversity conditions” which, together with a tractable optimism-based algorithm, guarantee sample-efficient recovery of this natural hierarchy. Furthermore, we provide regret bounds on a learner using the recovered hierarchy to solve a meta-test task. Our bounds incorporate common notions in HRL literature such as temporal and state/action abstractions, suggesting that our setting and analysis capture important features of HRL in practice.
****
Convolutional Neural Network Dynamics: A Graph Perspective
Fatemeh Vahedian,Ruiyu Li,Puja Trivedi,Di Jin,Danai Koutra
The success of neural networks (NNs) in a wide range of applications has led to increased interest in understanding the underlying learning dynamics of these models. In this paper, we go beyond mere descriptions of the learning dynamics by taking a graph perspective and investigating the relationship between the graph structure of NNs and their performance.
Specifically, we propose (1) representing the neural network learning process as a time-evolving graph (i.e., a series of static graph snapshots over epochs), (2) capturing the structural changes of the NN during the training phase in a simple temporal summary, and (3) leveraging the structural summary to predict the accuracy of the underlying NN in a classification or regression task. For the dynamic graph representation of NNs, we explore structural representations for fully-connected and convolutional layers, which are key components of powerful NN models. Our analysis shows that a simple summary of graph statistics, such as weighted degree and eigenvector centrality, over just a few epochs, can be used to accurately predict the performance of NNs. For example, a weighted degree-based summary of the time-evolving graph that is constructed based on 5 training epochs of the LeNet architecture achieves classification accuracy of over 93\%. Our findings are consistent for different NN architectures, including LeNet, VGG, AlexNet, and ResNet.
****
A General Analysis of Example-Selection for Stochastic Gradient Descent
Yucheng Lu,Si Yi Meng,Christopher De Sa
Training example order in SGD has long been known to affect convergence rate. Recent results show that accelerated rates are possible in a variety of cases for permutation-based sample orders, in which each example from the training set is used once before any example is reused. In this paper, we develop a broad condition on the sequence of examples used by SGD that is sufficient to prove tight convergence rates in both strongly convex and non-convex settings. We show that our approach suffices to recover, and in some cases improve upon, previous state-o

f-the-art analyses for four known example-selection schemes: (1) shuffle once, (2) random reshuffling, (3) random reshuffling with data echoing, and (4) Markov Chain Gradient Descent. Motivated by our theory, we propose two new example-selection approaches. First, using quasi-Monte-Carlo methods, we achieve unprecedented accelerated convergence rates for learning with data augmentation. Second, we greedily choose a fixed scan-order to minimize the metric used in our condition and show that we can obtain more accurate solutions from the same number of epochs of SGD. We conclude by empirically demonstrating the utility of our approach for both convex linear-model and deep learning tasks. Our code is available at: https://github.com/EugeneLYC/qmc-ordering.

****

GraphEBM: Towards Permutation Invariant and Multi-Objective Molecular Graph Generation

Meng Liu,Keqiang Yan,Bora Oztekin,Shuiwang Ji

Although significant progress has been made in molecular graph generation recently, permutation invariance and multi-objective generation remain to be important but challenging goals to achieve. In this work, we propose GraphEBM, a molecular graph generation method via energy-based models (EBMs), as an exploratory work to perform permutation invariant and multi-objective molecule generation. Particularly, thanks to the flexibility of EBMs and our parameterized permutation-invariant energy function, our GraphEBM can define a permutation invariant distribution over molecular graphs. We learn the energy function by contrastive divergence and generate samples by Langevin dynamics. In addition, to generate molecules with a specific desirable property, we propose a simple yet effective learning strategy, which pushes down energies with flexible degrees according to the properties of corresponding molecules. Further, we explore to use our GraphEBM for generating molecules towards multiple objectives via compositional generation, which is practically desired in drug discovery. We conduct comprehensive experiments on random, single-objective, and multi-objective molecule generation tasks. The results demonstrate our method is effective.

****

SimVLM: Simple Visual Language Model Pretraining with Weak Supervision

Zirui Wang,Jiahui Yu,Adams Wei Yu,Zihang Dai,Yulia Tsvetkov,Yuan Cao

With recent progress in joint modeling of visual and textual representations, Vision-Language Pretraining (VLP) has achieved impressive performance on many multimodal downstream tasks. However, the requirement for expensive annotations including clean image captions and regional labels limits the scalability of existing approaches, and complicates the pretraining procedure with the introduction of multiple dataset-specific objectives. In this work, we relax these constraints and present a minimalist pretraining framework, named Simple Visual Language Model (SimVLM). Unlike prior work, SimVLM reduces the training complexity by exploiting large-scale weak supervision, and is trained end-to-end with a single prefix language modeling objective. Without utilizing extra data or task-specific customization, the resulting model significantly outperforms previous pretraining methods and achieves new state-of-the-art results on a wide range of discriminative and generative vision-language benchmarks, including VQA (+3.74% vqa-score), NLVR2 (+1.17% accuracy), SNLI-VE (+1.37% accuracy) and image captioning tasks (+10.1% average CIDEr score). Furthermore, we demonstrate that SimVLM acquires strong generalization and transfer ability, enabling zero-shot behavior including open-ended visual question answering and cross-modality transfer.

****

Gradient-Guided Importance Sampling for Learning Discrete Energy-Based Models

Meng Liu,Haoran Liu,Shuiwang Ji

Learning energy-based models (EBMs) is known to be difficult especially on discrete data where gradient-based learning strategies cannot be applied directly. Although ratio matching is a sound method to learn discrete EBMs, it suffers from expensive computation and excessive memory requirement, thereby resulting in difficulties for learning EBMs on high-dimensional data. In this study, we propose ratio matching with gradient-guided importance sampling (RMwGGIS) to alleviate the above limitations. Particularly, we leverage the gradient of the energy funct ion w.r.t. the discrete data space to approximately construct the provable optimal proposal distribution, which is subsequently used by importance sampling to efficiently estimate the original ratio matching objective. We perform experiments on density modeling over synthetic discrete data and graph generation to evaluate our proposed method. The experimental results demonstrate that our method can significantly alleviate the limitations of ratio matching and perform more effectively in practice.

****

Bundle Networks: Fiber Bundles, Local Trivializations, and a Generative Approach to Exploring Many-to-one Maps

Nico Courts,Henry Kvinge

Many-to-one maps are ubiquitous in machine learning, from the image recognition model that assigns a multitude of distinct images to the concept of “cat” to the time series forecasting model which assigns a range of distinct time-series to a single scalar regression value. While the primary use of such models is naturally to associate correct output to each input, in many problems it is also useful to be able to explore, understand, and sample from a model's fibers, which are the set of input values x such that $f(x) = y,$ for fixed y in the output space. In this paper we show that popular generative architectures are ill-suited to such tasks. Motivated by this, we introduce a novel generative architecture, Bundle Networks, based on the concept of a fiber bundle from (differential) topology. BundleNets exploit the idea of a local trivialization wherein a space can be locally decomposed into a product space that cleanly encodes the many-to-one nature of the map. By enforcing this decomposition in BundleNets and by utilizing state-of-the-art invertible components, investigating a network's fibers becomes natural.

****

On Heterogeneously Distributed Data, Sparsity Matters

Tiansheng Huang,Shiwei Liu,Li Shen,Fengxiang He,Weiwei Lin,Dacheng Tao

Federated learning (FL) is particularly vulnerable to heterogeneously distributed data, since a common global model in FL may not adapt to the heterogeneous data distribution of each user. To counter this issue, personalized FL (PFL) was proposed to produce dedicated local models for each individual user. However, PFL is far from its maturity, because existing PFL solutions either demonstrate unsatisfactory generalization towards different model architectures or cost enormous extra computation and memory. In this work, we propose federated learning with personalized sparse mask (FedSpa), a novel personalized federated learning scheme that employs personalized sparse masks to customize sparse local models on the edge. Instead of training fully dense PFL models, FedSpa only maintains a fixed number of active parameters throughout training (aka sparse-to-sparse training), which enables users' models to achieve personalization with consistently cheap communication, computation, and memory cost. We theoretically show that with the rise of data heterogeneity, setting a higher sparsity of FedSpa may potentially result in a smaller error bound on its personalized models, which also coincides with our empirical observations. Comprehensive experiments demonstrate that FedSpa significantly saves communication and computation costs, while simultaneously achieves higher model accuracy and faster convergence speed against several state-of-the-art PFL methods.

****

Privacy Implications of Shuffling

Casey Meehan,Amrita Roy Chowdhury,Kamalika Chaudhuri,Somesh Jha

\ldp deployments are vulnerable to inference attacks as an adversary can link the noisy responses to their identity and subsequently, auxiliary information using the \textit{order} of the data. An alternative model, shuffle \textsf{DP}, prevents this by shuffling the noisy responses uniformly at random. However, this limits the data learnability -- only symmetric functions (input order agnostic) can be learned. In this paper, we strike a balance and show that systematic shuffling of the noisy responses can thwart specific inference attacks while retaining some meaningful data learnability. To this end, we propose a novel privacy guarantee, \name-privacy, that captures the privacy of the order of a data sequenc

e. \name-privacy allows tuning the granularity at which the ordinal information is maintained, which formalizes the degree the resistance to inference attacks trading it off with data learnability. Additionally, we propose a novel shuffling mechanism that can achieve \name-privacy and demonstrate the practicality of our mechanism via evaluation on real-world datasets.

****

DL-based prediction of optimal actions of human experts

Jung H. Lee,Ryan S Butner,Elise Saxon,Nathan Oken Hodas

Expert systems have been developed to emulate human experts’ decision-making. Once developed properly, expert systems can assist or substitute human experts, but they require overly expensive knowledge engineering/acquisition. Notably, deep learning (DL) can train highly efficient computer vision systems only from examples instead of relying on carefully selected feature sets by human experts. Thus, we hypothesize that DL can be used to build expert systems that can learn human experts’ decision-making from examples only without relying on overly expensive knowledge engineering. To address this hypothesis, we train DL agents to predict optimal strategies (actions or action sequences) for the popular game `Angry Birds’, which requires complex problem-solving skills. In our experiments, after being trained with screenshots of different levels and pertinent 3-star guides, DL agents can predict strategies for unseen levels. This raises the possibility of building DL-based expert systems that do not require overly expensive knowledge engineering.

****

HALP: Hardware-Aware Latency Pruning

Maying Shen,Hongxu Yin,Pavlo Molchanov,Lei Mao,Jianna Liu,Jose M. Alvarez

Structural pruning can simplify network architecture and improve the inference speed. We propose Hardware-Aware Latency Pruning (HALP) 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, HALP leverages latency lookup table to track latency reduction potential and global saliency score to gauge on 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 HALP to surpass prior work in pruning efficacy and accuracy-efficiency trade-off. We examine HALP on both classification and detection tasks, over varying networks, on ImageNet1K and VOC datasets. In particular for ResNet-50/-101 pruning on ImageNet1K, HALP improves network speed by $1.60\times$/$1.90\times$ with $+0.3\%$/-0.2% top-1 accuracy changes, respectively. For SSD pruning on VOC, HALP improves throughput by $1.94\times$ with only a 0.56 mAP drop. HALP consistently outperforms prior art, sometimes by large margins.

****

On the role of population heterogeneity in emergent communication

Mathieu Rita,Florian Strub,Jean-Bastien Grill,Olivier Pietquin,Emmanuel Dupoux

Populations have often been perceived as a structuring component for language to emerge and evolve: the larger the population, the more systematic the language. While this observation is widespread in the sociolinguistic literature, it has not been reproduced in computer simulations with neural agents. In this paper, we thus aim to clarify this apparent contradiction. We explore emergent language properties by varying agent population size in the speaker-listener Lewis Game. After reproducing the experimental paradox, we challenge the simulation assumption that the agent community is homogeneous. We first investigate how speaker-listener asymmetry alters language structure to examine two potential diversity factors: training speed and network capacity. We find out that emergent language properties are only altered by the relative difference of factors between speaker and listener, and not by their absolute values. From then, we leverage this observation to control population heterogeneity without introducing confounding factors. We finally show that introducing such training speed heterogeneities naturally sort out the initial paradox: larger simulated communities start developing more systematic and structured languages.

****

Hindsight is 20/20: Leveraging Past Traversals to Aid 3D Perception

Yurong You,Katie Z Luo,Xiangyu Chen,Junan Chen,Wei-Lun Chao,Wen Sun,Bharath Hariharan,Mark Campbell,Kilian Q Weinberger

Self-driving cars must detect vehicles, pedestrians, and other traffic participants accurately to operate safely. Small, far-away, or highly occluded objects are particularly challenging because there is limited information in the LiDAR point clouds for detecting them. To address this challenge, we leverage valuable information from the past: in particular, data collected in past traversals of the same scene. We posit that these past data, which are typically discarded, provide rich contextual information for disambiguating the above-mentioned challenging cases. To this end, we propose a novel end-to-end trainable Hindsight framework to extract this contextual information from past traversals and store it in an easy-to-query data structure, which can then be leveraged to aid future 3D object detection of the same scene. We show that this framework is compatible with most modern 3D detection architectures and can substantially improve their average precision on multiple autonomous driving datasets, most notably by more than 300% on the challenging cases. Our code is available at https://github.com/YurongYou/Hindsight.

****

Analogies and Feature Attributions for Model Agnostic Explanation of Similarity Learners

Karthikeyan Natesan Ramamurthy,Amit Dhurandhar,Dennis Wei,Zaid Bin Tariq

Post-hoc explanations for black box models have been studied extensively in classification and regression settings. However, explanations for models that output similarity between two inputs have received comparatively lesser attention. In this paper, we provide model agnostic local explanations for similarity learners applicable to tabular and text data. We first propose a method that provides feature attributions to explain the similarity between a pair of inputs as determined by a black box similarity learner. We then propose analogies as a new form of explanation in machine learning. Here the goal is to identify diverse analogous pairs of examples that share the same level of similarity as the input pair and provide insight into (latent) factors underlying the model's prediction. The selection of analogies can optionally leverage feature attributions, thus connecting the two forms of explanation while still maintaining complementarity. We prove that our analogy objective function is submodular, making the search for good-quality analogies efficient. We apply the proposed approaches to explain similarities between sentences as predicted by a state-of-the-art sentence encoder, and between patients in a healthcare utilization application. Efficacy is measured through quantitative evaluations, a careful user study, and examples of explanations.

****

AutoNF: Automated Architecture Optimization of Normalizing Flows Using a Mixture Distribution Formulation

Yu Wang,Jan Drgona,Jiaxin Zhang,Karthik Somayaji NS,Frank Y Liu,Malachi Schram,Peng Li

Although various flow models based on different transformations have been proposed, there still lacks a quantitative analysis of performance-cost trade-offs between different flows as well as a systematic way of constructing the best flow architecture. To tackle this challenge, we present an automated normalizing flow (NF) architecture search method. Our method aims to find the optimal sequence of transformation layers from a given set of unique transformations with three folds. First, a mixed distribution is formulated to enable efficient architecture optimization originally on the discrete space without violating the invertibility of the resulting NF architecture. Second, the mixture NF is optimized with an approximate upper bound which has a more preferable global minimum. Third, a block-wise alternating optimization algorithm is proposed to ensure efficient architecture optimization of deep flow models.

****

Language-driven Semantic Segmentation

Boyi Li,Kilian Q Weinberger,Serge Belongie,Vladlen Koltun,Rene Ranftl

We present LSeg, a novel model for language-driven semantic image segmentation. LSeg uses a text encoder to compute embeddings of descriptive input labels (e.g., ``grass'' or ``building'') together with a transformer-based image encoder that computes dense per-pixel embeddings of the input image. The image encoder is trained with a contrastive objective to align pixel embeddings to the text embedding of the corresponding semantic class. The text embeddings provide a flexible label representation in which semantically similar labels map to similar regions in the embedding space (e.g., ``cat'' and ``furry''). This allows LSeg to generalize to previously unseen categories at test time, without retraining or even requiring a single additional training sample. We demonstrate that our approach achieves highly competitive zero-shot performance compared to existing zero- and few-shot semantic segmentation methods, and even matches the accuracy of traditional segmentation algorithms when a fixed label set is provided. Code and demo are available at https://github.com/isl-org/lang-seg.

****

Image BERT Pre-training with Online Tokenizer

Jinghao Zhou,Chen Wei,Huiyu Wang,Wei Shen,Cihang Xie,Alan Yuille,Tao Kong

The success of language Transformers is primarily attributed to the pretext task of masked language modeling (MLM), where texts are first tokenized into semantically meaningful pieces.

In this work, we study masked image modeling (MIM) and indicate the necessity and challenges of using a semantically meaningful visual tokenizer.

We present a self-supervised framework iBOT that can perform masked prediction with an online tokenizer.

Specifically, we perform self-distillation on masked patch tokens and take the teacher network as the online tokenizer, along with self-distillation on the class token to acquire visual semantics.

The online tokenizer is jointly learnable with the MIM objective and dispenses with a multi-stage training pipeline where the tokenizer needs to be pre-trained beforehand.

We show the prominence of iBOT by achieving an 82.3% linear probing accuracy and an 87.8% fine-tuning accuracy evaluated on ImageNet-1K.

Beyond the state-of-the-art image classification results, we underline emerging local semantic patterns, which helps the models to obtain strong robustness against common corruptions and achieve leading results on dense downstream tasks, e.g., object detection, instance segmentation, and semantic segmentation.

****

Assessing Generalization of SGD via Disagreement

Yiding Jiang,Vaishnavh Nagarajan,Christina Baek,J Zico Kolter

We empirically show that the test error of deep networks can be estimated by training the same architecture on the same training set but with two different runs of Stochastic Gradient Descent (SGD), and then measuring the disagreement rate between the two networks on unlabeled test data. This builds on -- and is a stronger version of -- the observation in Nakkiran&Bansal 20, which requires the runs to be on separate training sets. We further theoretically show that this peculiar phenomenon arises from the well-calibrated nature of ensembles of SGD-trained models. This finding not only provides a simple empirical measure to directly predict the test error using unlabeled test data, but also establishes a new conceptual connection between generalization and calibration.

****

On Learning to Solve Cardinality Constrained Combinatorial Optimization in One-Shot: A Re-parameterization Approach via Gumbel-Sinkhorn-TopK

Runzhong Wang,Li Shen,Yiting Chen,Junchi Yan,Xiaokang Yang,Dacheng Tao

Cardinality constrained combinatorial optimization requires selecting an optimal subset of k elements, and it will be appealing to design data-driven algorithms that perform TopK selection over a probability distribution predicted by a neural network. However, the existing differentiable TopK operator suffers from an unbounded gap between the soft prediction and the discrete solution, leading to inaccurate estimation of the combinatorial objective score. In this paper, we p resent a self-supervised learning pipeline for cardinality constrained combinatorial optimization, which incorporates with Gumbel-Sinkhorn-TopK (GS-TopK) for near-discrete TopK predictions and the re-parameterization trick resolving the non-differentiable challenge. Theoretically, we characterize a bounded gap between the Maximum-A-Posteriori (MAP) inference and our proposed method, resolving the divergence issue in the previous differentiable TopK operator and also providing a more accurate estimation of the objective score given a provable tightened bound to the discrete decision variables. Experiments on max covering and discrete clustering problems show that our method outperforms state-of-the-art Gurobi solver and the novel one-shot learning method Erdos Goes Neural.

****

Robust Deep Neural Networks for Heterogeneous Tabular Data

Vadim Borisov,Klaus Broelemann,Enkelejda Kasneci,Gjergji. Kasneci

Although deep neural networks (DNNs) constitute the state-of-the-art in many tasks based on image, audio, or text data, their performance on heterogeneous, tabular data is typically inferior to that of decision tree ensembles. To bridge the gap between the difficulty of DNNs to handle tabular data and leverage the flexibility of deep learning under input heterogeneity, we propose DeepTLF, a framework for deep tabular learning. The core idea of our method is to transform the heterogeneous input data into homogeneous data to boost the performance of DNNs considerably. For the transformation step, we develop a novel knowledge distillations approach, TreeDrivenEncoder, which exploits the structure of decision trees trained on the available heterogeneous data to map the original input vectors onto homogeneous vectors that a DNN can use to improve the predictive performance. Through extensive and challenging experiments on various real-world datasets, we demonstrate that the DeepTLF pipeline leads to higher predictive performance. On average, our framework shows 19.6\% performance improvement in comparison to DNNs. The DeepTLF code is publicly available.

****

Locally Invariant Explanations: Towards Causal Explanations through Local Invariant Learning

Amit Dhurandhar,Karthikeyan Natesan Ramamurthy,Kartik Ahuja,Vijay Arya

Locally interpretable model agnostic explanations (LIME) method is one of the most popular methods used to explain black-box models at a per example level. Although many variants have been proposed, few provide a simple way to produce high fidelity explanations that are also stable and intuitive. In this work, we provide a novel perspective by proposing a model agnostic local explanation method inspired by the invariant risk minimization (IRM) principle -- originally proposed for (global) out-of-distribution generalization -- to provide such high fidelity explanations that are also stable and unidirectional across nearby examples. Our method is based on a game theoretic formulation where we theoretically show that our approach has a strong tendency to eliminate features where the gradient of the black-box function abruptly changes sign in the locality of the example we want to explain, while in other cases it is more careful and will choose a more conservative (feature) attribution, a behavior which can be highly desirable for recourse. Empirically, we show on tabular, image and text data that the quality of our explanations with neighborhoods formed using random perturbations are much better than LIME and in some cases even comparable to other methods that use realistic neighbors sampled from the data manifold, where the latter is a popular strategy to obtain high quality explanations. This is a desirable property given that learning a manifold to either create realistic neighbors or to project explanations is typically expensive or may even be impossible. Moreover, our algorithm is simple and efficient to train, and can ascertain stable input features for local decisions of a black-box without access to side information such as a (partial) causal graph as has been seen in some recent works.

****

ModeRNN: Harnessing Spatiotemporal Mode Collapse in Unsupervised Predictive Learning

Zhiyu Yao,Yunbo Wang,Haixu Wu,Jianmin Wang,Mingsheng Long

Learning predictive models for unlabeled spatiotemporal data is challenging in part because visual dynamics can be highly entangled in real scenes, making existing approaches prone to overfit partial modes of physical processes while neglecting to reason about others. We name this phenomenon \textit{spatiotemporal mode collapse} and explore it for the first time in predictive learning. The key is to provide the model with a strong inductive bias to discover the compositional structures of latent modes. To this end, we propose ModeRNN, which introduces a novel method to learn structured hidden representations between recurrent states. The core idea of this framework is to first extract various components of visual dynamics using a set of \textit{spatiotemporal slots} with independent parameters. Considering that multiple space-time patterns may co-exist in a sequence, we leverage learnable importance weights to adaptively aggregate slot features into a unified hidden representation, which is then used to update the recurrent states. Across the entire dataset, different modes result in different responses on the mixtures of slots, which enhances the ability of ModeRNN to build structured representations and thus prevents the so-called mode collapse. Unlike existing models, ModeRNN is shown to prevent spatiotemporal mode collapse and further benefit from learning mixed visual dynamics.

****

Generative Planning for Temporally Coordinated Exploration in Reinforcement Learning

Haichao Zhang,Wei Xu,Haonan Yu

Standard model-free reinforcement learning algorithms optimize a policy that generates the action to be taken in the current time step in order to maximize expected future return. While flexible, it faces difficulties arising from the inefficient exploration due to its single step nature. In this work, we present Generative Planning method (GPM), which can generate actions not only for the current step, but also for a number of future steps (thus termed as generative planning). This brings several benefits to GPM. Firstly, since GPM is trained by maximizing value, the plans generated from it can be regarded as intentional action sequences for reaching high value regions. GPM can therefore leverage its generated multi-step plans for temporally coordinated exploration towards high value regions, which is potentially more effective than a sequence of actions generated by perturbing each action at single step level, whose consistent movement decays exponentially with the number of exploration steps. Secondly, starting from a crude initial plan generator, GPM can refine it to be adaptive to the task, which, in return, benefits future explorations. This is potentially more effective than commonly used action-repeat strategy, which is non-adaptive in its form of plans. Additionally, since the multi-step plan can be interpreted as the intent of the agent from now to a span of time period into the future, it offers a more informative and intuitive signal for interpretation. Experiments are conducted on several benchmark environments and the results demonstrated its effectiveness compared with several baseline methods.

****

The Remarkable Effectiveness of Combining Policy and Value Networks in A*-based Deep RL for AI Planning

Dieqiao Feng,Carla P Gomes,Bart Selman

Despite the tremendous success of applying traditional backtrack-style combinatorial search methods in various NP-complete domains such as SAT and CSP as well as using deep reinforcement learning (RL) to tackle two-player games such as Go, PSPACE-hard AI planning has remained out of reach for current AI planning systems. Even carefully designed domain-specific solvers fail quickly due to the exponential combinatorial search space on hard instances. Recent work based on deep learning guided search algorithms that combine traditional search-based methods, such as A\textsc{*} and MCTS search, with deep neural networks' heuristic prediction has shown promising progress. These methods can solve a significant number of hard planning instances beyond specialized solvers. To better understanding why these approaches work we study the interplay of the policy and value networks in A\textsc{*}-based deep RL and show the surprising effectiveness of the policy network, further enhanced by the value network, as a guiding heuristic for A\

textsc{*}. To further understand the phenomena, we study the cost distributions of deep planners and found planning instances can have heavy-tailed runtime distributions, with tails both on the right-hand and left-hand sides. In particular, for the first time, we show the existence of {\textit{left}} heavy tails and propose a theoretical model that can explain the appearance of these tails. We provide extensive experimental data supporting our model. The experiments show the critical role of the policy network as a powerful heuristic guiding A\textsc{*}, which can lead to left tails with polynomial scaling by avoiding exploring exponential size sub-trees early on in the search. Our results also demonstrate the importance of random restart strategies, as are widely used in traditional combinatorial solvers, for deep reinforcement learning and deep AI planning systems to avoid left and right heavy tails.

****

Accelerated Policy Learning with Parallel Differentiable Simulation

Jie Xu,Viktor Makoviychuk,Yashraj Narang,Fabio Ramos,Wojciech Matusik,Animesh Garg,Miles Macklin

Deep reinforcement learning can generate complex control policies, but requires large amounts of training data to work effectively. Recent work has attempted to address this issue by leveraging differentiable simulators. However, inherent problems such as local minima and exploding/vanishing numerical gradients prevent these methods from being generally applied to control tasks with complex contact-rich dynamics, such as humanoid locomotion in classical RL benchmarks. In this work we present a high-performance differentiable simulator and a new policy learning algorithm (SHAC) that can effectively leverage simulation gradients, even in the presence of non-smoothness. Our learning algorithm alleviates problems with local minima through a smooth critic function, avoids vanishing/exploding gradients through a truncated learning window, and allows many physical environments to be run in parallel. We evaluate our method on classical RL control tasks, and show substantial improvements in sample efficiency and wall-clock time over state-of-the-art RL and differentiable simulation-based algorithms. In addition, we demonstrate the scalability of our method by applying it to the challenging high-dimensional problem of muscle-actuated locomotion with a large action space, achieving a greater than $17\times$ reduction in training time over the best-performing established RL algorithm. More visual results are provided at: https://short-horizon-actor-critic.github.io/.

****

Generative Negative Replay for Continual Learning

Gabriele Graffieti,Davide Maltoni,Lorenzo Pellegrini,Vincenzo Lomonaco

Learning continually is a key aspect of intelligence and a necessary ability to solve many real-world problems. One of the most effective strategies to control catastrophic forgetting, the Achilles’ heel of continual learning, is storing part of the old data and replay them interleaved with new experiences (also known as the replay approach). Generative replay, that is using generative models to provide replay patterns on demand, is particularly intriguing, however, it was shown to be effective mainly under simplified assumptions, such as simple scenarios and low-dimensional benchmarks.

In this paper, we show that, while the generated data are usually not able to improve the classification accuracy for the old classes, they can be effective as negative examples (or antagonists) to learn the new classes, especially when the learning experiences are small and contain examples of just one or few classes. The proposed approach is validated on complex class-incremental and data-incremental continual learning scenarios (CORe50 and ImageNet-1000) composed of high-dimensional data and a large number of training experiences: a setup where existing generative replay approaches usually fail.

****

Do We Need Anisotropic Graph Neural Networks?

Shyam A. Tailor,Felix Opolka,Pietro Lio,Nicholas Donald Lane

Common wisdom in the graph neural network (GNN) community dictates that anisotropic models---in which messages sent between nodes are a function of both the source and target node---are required to achieve state-of-the-art performance. Benc hmarks to date have demonstrated that these models perform better than comparable isotropic models---where messages are a function of the source node only. In this work we provide empirical evidence challenging this narrative: we propose an isotropic GNN, which we call Efficient Graph Convolution (EGC), that consistently outperforms comparable anisotropic models, including the popular GAT or PNA architectures by using spatially-varying adaptive filters. In addition to raising important questions for the GNN community, our work has significant real-world implications for efficiency. EGC achieves higher model accuracy, with lower memory consumption and latency, along with characteristics suited to accelerator implementation, while being a drop-in replacement for existing architectures. As an isotropic model, it requires memory proportional to the number of vertices in the graph ($\mathcal{O}(V)$); in contrast, anisotropic models require memory proportional to the number of edges ($\mathcal{O}(E)$). We demonstrate that EGC outperforms existing approaches across 6 large and diverse benchmark datasets, and conclude by discussing questions that our work raise for the community going forward. Code and pretrained models for our experiments are provided at https://github.com/shyam196/egc.

****

Is High Variance Unavoidable in RL? A Case Study in Continuous Control
Johan Bjorck,Carla P Gomes,Kilian Q Weinberger

Reinforcement learning (RL) experiments have notoriously high variance, and minor details can have disproportionately large effects on measured outcomes. This is problematic for creating reproducible research and also serves as an obstacle when applying RL to sensitive real-world applications. In this paper, we investigate causes for this perceived instability. To allow for an in-depth analysis, we focus on a specifically popular setup with high variance -- continuous control from pixels with an actor-critic agent. In this setting, we demonstrate that poor outlier runs which completely fail to learn are an important source of variance, but that weight initialization and initial exploration are not at fault. We show that one cause for these outliers is unstable network parametrization which leads to saturating nonlinearities. We investigate several fixes to this issue and find that simply normalizing penultimate features is surprisingly effective. For sparse tasks, we also find that partially disabling clipped double Q-learning decreases variance. By combining fixes we significantly decrease variances, lowering the average standard deviation across 21 tasks by a factor >3 for a state-of-the-art agent. This demonstrates that the perceived variance is not necessarily inherent to RL. Instead, it may be addressed via simple modifications and we argue that developing low-variance agents is an important goal for the RL community.

****

What Would the Expert $do(\cdot)$?: Causal Imitation Learning
Gokul Swamy,Sanjiban Choudhury,Drew Bagnell,Steven Wu

We develop algorithms for imitation learning from data that was corrupted by unobserved confounders. Sources of such confounding include (a) persistent perturbations to actions or (b) the expert responding to a part of the state that the learner does not have access to. When a confounder affects multiple timesteps of recorded data, it can manifest as spurious correlations between states and actions that a learner might latch onto, leading to poor policy performance. By utilizing the effect of past states on current states, we are able to break up these spurious correlations, an application of the econometric technique of instrumental variable regression. This insight leads to two novel algorithms, one of a generative-modeling flavor ($\texttt{DoubIL}$) that can utilize access to a simulator and one of a game-theoretic flavor ($\texttt{ResiduIL}$) that can be run offline. Both approaches are able to find policies that match the result of a query to an unconfounded expert. We find both algorithms compare favorably to non-causal approaches on simulated control problems.

****

Sparsistent Model Discovery
Georges Tod,Gert-Jan Both,Remy Kusters

Discovering the partial differential equations underlying spatio-temporal datase

ts from very limited and highly noisy observations is of paramount interest in many scientific fields. However, it remains an open question to know when model discovery algorithms based on sparse regression can actually recover the underlying physical processes. In this work, we show the design matrices used to infer the equations by sparse regression can violate the irrepresentability condition (IRC) of the Lasso, even when derived from analytical PDE solutions (i.e. without additional noise). Sparse regression techniques which can recover the true underlying model under violated IRC conditions are therefore required, leading to the introduction of the randomised adaptive Lasso. We show once the latter is integrated within the deep learning model discovery framework DeepMod, a wide variety of nonlinear and chaotic canonical PDEs can be recovered: (1) up to $\mathcal{O}(2)$ higher noise-to-sample ratios than state-of-the-art algorithms, (2) with a single set of hyperparameters, which paves the road towards truly automated model discovery.
****

Simple GNN Regularisation for 3D Molecular Property Prediction and Beyond

Jonathan Godwin,Michael Schaarschmidt,Alexander L Gaunt,Alvaro Sanchez-Gonzalez,Yulia Rubanova,Petar Veli■kovi■,James Kirkpatrick,Peter Battaglia

In this paper we show that simple noisy regularisation can be an effective way to address oversmoothing. We first argue that regularisers ad-dressing oversmoothing should both penalise node latent similarity and encourage meaningful node representations. From this observation we derive “Noisy Nodes”,a simple technique in which we corrupt the input graph with noise, and add a noise correcting node-level loss. The diverse node level loss encourages latent node diversity, and the denoising objective encourages graph manifold learning. Our regulariser applies well-studied methods in simple, straightforward ways which allow even generic architectures to overcome oversmoothing and achieve state of the art results on quantum chemistry tasks such as QM9 and Open Catalyst, and improve results significantly on Open Graph Benchmark (OGB) datasets. Our results suggest Noisy Nodes can serve as a complementary building block in the GNN toolkit.
****

Should We Be Pre-training? An Argument for End-task Aware Training as an Alternative

Lucio M. Dery,Paul Michel,Ameet Talwalkar,Graham Neubig

In most settings of practical concern, machine learning practitioners know in advance what end-task they wish to boost with auxiliary tasks. However, widely used methods for leveraging auxiliary data like pre-training and its continued-pretraining variant are end-task agnostic: they rarely, if ever, exploit knowledge of the target task. We study replacing end-task agnostic continued training of pre-trained language models with end-task aware training of said models. We argue that for sufficiently important end-tasks, the benefits of leveraging auxiliary data in a task-aware fashion can justify forgoing the traditional approach of obtaining generic, end-task agnostic representations as with (continued) pre-training. On three different low-resource NLP tasks from two domains, we demonstrate that multi-tasking the end-task and auxiliary objectives results in significantly better downstream task performance than the widely-used task-agnostic continued pre-training paradigm of Gururangan et al. (2020).

We next introduce an online meta-learning algorithm that learns a set of multi-task weights to better balance among our multiple auxiliary objectives, achieving further improvements on end-task performance and data efficiency.
****

Composing Partial Differential Equations with Physics-Aware Neural Networks

Matthias Karlbauer,Timothy Praditia,Sebastian Otte,Sergey Oladyshkin,Wolfgang Nowak,Martin V. Butz

We introduce a compositional physics-aware neural network (FINN) for learning spatiotemporal advection-diffusion processes. FINN implements a new way of combining the learning abilities of artificial neural networks with physical and structural knowledge from numerical simulation by modeling the constituents of partial differential equations (PDEs) in a compositional manner. Results on both one- and two-dimensional PDEs (Burger's, diffusion-sorption, diffusion-reaction) demon

strate FINN's superior process modeling accuracy and excellent out-of-distribution generalization ability beyond initial and boundary conditions. With only one tenth of the number of parameters on average, FINN outperforms pure machine learning and other state-of-the-art physics-aware models in all cases---often even by multiple orders of magnitude. Moreover, FINN outperforms a calibrated physical model when approximating sparse real-world data in a diffusion-sorption scenario, confirming its generalization abilities and showing explanatory potential by revealing the unknown retardation factor of the observed process.
****

Counterfactual Graph Learning for Link Prediction
Tong Zhao,Gang Liu,Daheng Wang,Wenhao Yu,Meng Jiang
Learning to predict missing links is important for many graph-based applications. Existing methods were designed to learn the association between two sets of variables: (1) the observed graph structure (e.g., clustering effect) and (2) the existence of link between a pair of nodes. However, the causal relationship between these variables was ignored. We visit the possibility of learning it by asking a counterfactual question: “would the link exist or not if the observed graph structure became different?” To answer this question, we leverage causal models considering the information of the node pair (i.e., learned graph representations) as context, global graph structural properties as treatment, and link existence as outcome. In this work, we propose a novel link prediction method that enhances graph learning by counterfactual inference. It creates counterfactual links from the observed ones, and learns representations from both the observed and counterfactual links. Experiments on benchmark datasets show that this novel graph learning method achieves state-of-the-art performance on link prediction.
****

DeepSplit: Scalable Verification of Deep Neural Networks via Operator Splitting
Shaoru Chen,Eric Wong,J Zico Kolter,Mahyar Fazlyab
Analyzing the worst-case performance of deep neural networks against input perturbations amounts to solving a large-scale non-convex optimization problem, for which several past works have proposed convex relaxations as a promising alternative. However, even for reasonably-sized neural networks, these relaxations are not tractable, and so must be replaced by even weaker relaxations in practice. In this work, we propose a novel operator splitting method that can directly solve a convex relaxation of the problem to high accuracy, by splitting it into smaller sub-problems that often have analytical solutions. The method is modular, scales to very large problem instances, and compromises of operations that are amenable to fast parallelization with GPU acceleration. We demonstrate our method in obtaining tighter bounds on the worst-case performance of large convolutional networks in image classification and reinforcement learning settings.
****

Positive-Unlabeled Learning with Uncertainty-aware Pseudo-label Selection
Emilio Dorigatti,Jann Goschenhofer,Benjamin Schubert,Mina Rezaei,Bernd Bischl
Positive-unlabeled (PU) learning aims at learning a binary classifier from only positive and unlabeled training data. Recent approaches address this problem via cost-sensitive learning by developing unbiased loss functions or via iterative pseudo-labeling solutions to further improve performance. However, two-steps procedures are vulnerable to incorrectly estimated pseudo-labels, as errors are propagated in later iterations when a new model is trained on erroneous predictions. To mitigate this issue we propose \textit{PUUPL}, a new loss-agnostic training procedure for PU learning that incorporates epistemic uncertainty in pseudo-labeling. Using an ensemble of neural networks and assigning pseudo-labels based on high confidence predictions improves the reliability of pseudo-labels, increasing the predictive performance of our method and leads to new state-of-the-art results in PU learning. With extensive experiments, we show the effectiveness of our method over different datasets, modalities, and learning tasks, as well as improved robustness over mispecifications of hyper-parameters and biased positive data. The source code of the method and all the experiments are available in the supplementary material.
****

Learning Super-Features for Image Retrieval

Philippe Weinzaepfel,Thomas Lucas,Diane Larlus,Yannis Kalantidis

Methods that combine local and global features have recently shown excellent performance on multiple challenging deep image retrieval benchmarks, but their use of local features raises at least two issues. First, these local features simply boil down to the localized map activations of a neural network, and hence can be extremely redundant. Second, they are typically trained with a global loss that only acts on top of an aggregation of local features; by contrast, testing is based on local feature matching, which creates a discrepancy between training and testing. In this paper, we propose a novel architecture for deep image retrieval, based solely on mid-level features that we call Super-features. These Super-features are constructed by an iterative attention module and constitute an ordered set in which each element focuses on a localized and discriminant image pattern. For training, they require only image labels. A contrastive loss operates directly at the level of Super-features and focuses on those that match across images. A second complementary loss encourages diversity. Experiments on common landmark retrieval benchmarks validate that Super-features substantially outperform state-of-the-art methods when using the same number of features, and only require a significantly smaller memory footprint to match their performance. Code and models are available at: https://github.com/naver/FIRe.

****

Pessimistic Bootstrapping for Uncertainty-Driven Offline Reinforcement Learning

Chenjia Bai,Lingxiao Wang,Zhuoran Yang,Zhi-Hong Deng,Animesh Garg,Peng Liu,Zhaoran Wang

Offline Reinforcement Learning (RL) aims to learn policies from previously collected datasets without exploring the environment. Directly applying off-policy algorithms to offline RL usually fails due to the extrapolation error caused by the out-of-distribution (OOD) actions. Previous methods tackle such problem by penalizing the Q-values of OOD actions or constraining the trained policy to be close to the behavior policy. Nevertheless, such methods typically prevent the generalization of value functions beyond the offline data and also lack precise characterization of OOD data. In this paper, we propose Pessimistic Bootstrapping for offline RL (PBRL), a purely uncertainty-driven offline algorithm without explicit policy constraints. Specifically, PBRL conducts uncertainty quantification via the disagreement of bootstrapped Q-functions, and performs pessimistic updates by penalizing the value function based on the estimated uncertainty. To tackle the extrapolating error, we further propose a novel OOD sampling method. We show that such OOD sampling and pessimistic bootstrapping yields provable uncertainty quantifier in linear MDPs, thus providing the theoretical underpinning for PBRL. Extensive experiments on D4RL benchmark show that PBRL has better performance compared to the state-of-the-art algorithms.

****

Equivariant Subgraph Aggregation Networks

Beatrice Bevilacqua,Fabrizio Frasca,Derek Lim,Balasubramaniam Srinivasan,Chen Cai,Gopinath Balamurugan,Michael M. Bronstein,Haggai Maron

Message-passing neural networks (MPNNs) are the leading architecture for deep learning on graph-structured data, in large part due to their simplicity and scalability. Unfortunately, it was shown that these architectures are limited in their expressive power. This paper proposes a novel framework called Equivariant Subgraph Aggregation Networks (ESAN) to address this issue. Our main observation is that while two graphs may not be distinguishable by an MPNN, they often contain distinguishable subgraphs. Thus, we propose to represent each graph as a set of subgraphs derived by some predefined policy, and to process it using a suitable equivariant architecture. We develop novel variants of the 1-dimensional Weisfeiler-Leman (1-WL) test for graph isomorphism, and prove lower bounds on the expressiveness of ESAN in terms of these new WL variants. We further prove that our approach increases the expressive power of both MPNNs and more expressive architectures. Moreover, we provide theoretical results that describe how design choices such as the subgraph selection policy and equivariant neural architecture affect our architecture's expressive power. To deal with the increased computationa l cost, we propose a subgraph sampling scheme, which can be viewed as a stochastic version of our framework. A comprehensive set of experiments on real and synthetic datasets demonstrates that our framework improves the expressive power and overall performance of popular GNN architectures.

****

Sparse Communication via Mixed Distributions

António Farinhas,Wilker Aziz,Vlad Niculae,Andre Martins

Neural networks and other machine learning models compute continuous representations, while humans communicate mostly through discrete symbols. Reconciling these two forms of communication is desirable for generating human-readable interpretations or learning discrete latent variable models, while maintaining end-to-end differentiability. Some existing approaches (such as the Gumbel-Softmax transformation) build continuous relaxations that are discrete approximations in the zero-temperature limit, while others (such as sparsemax transformations and the Hard Concrete distribution) produce discrete/continuous hybrids. In this paper, we build rigorous theoretical foundations for these hybrids, which we call "mixed random variables.'' Our starting point is a new "direct sum'' base measure defined on the face lattice of the probability simplex. From this measure, we introduce new entropy and Kullback-Leibler divergence functions that subsume the discrete and differential cases and have interpretations in terms of code optimality. Our framework suggests two strategies for representing and sampling mixed random variables, an extrinsic ("sample-and-project'’) and an intrinsic one (based on face stratification). We experiment with both approaches on an emergent communication benchmark and on modeling MNIST and Fashion-MNIST data with variational auto-encoders with mixed latent variables.

****

Tell me why!—Explanations support learning relational and causal structure

Andrew Kyle Lampinen,Nicholas Andrew Roy,Ishita Dasgupta,Stephanie C.Y. Chan,Allison Tam,Chen Yan,Adam Santoro,Neil Charles Rabinowitz,Jane X Wang,Felix Hill

Explanations play a considerable role in human learning, especially in areas that remain major challenges for AI—forming abstractions, and learning about the relational and causal structure of the world. Here, we explore whether machine learning models might likewise benefit from explanations. We outline a family of relational tasks that involve selecting an object that is the odd one out in a set (i.e., unique along one of many possible feature dimensions). Odd-one-out tasks require agents to reason over multi-dimensional relationships among a set of objects. We show that agents do not learn these tasks well from reward alone, but achieve >90% performance when they are also trained to generate language explaining object properties or why a choice is correct or incorrect. In further experiments, we show how predicting explanations enables agents to generalize appropriately from ambiguous, causally-confounded training, and even to meta-learn to perform experimental interventions to identify causal structure. We show that explanations help overcome the tendency of agents to fixate on simple features, and explore which aspects of explanations make them most beneficial. Our results suggest that learning from explanations is a powerful principle that could offer a promising path towards training more robust and general machine learning systems.

****

An Investigation on Hardware-Aware Vision Transformer Scaling

Chaojian Li,Kyungmin Kim,Bichen Wu,Peizhao Zhang,Hang Zhang,Xiaoliang Dai,Peter Vajda,Yingyan Lin

Vision Transformer (ViT) has demonstrated promising performance in various computer vision tasks, and recently attracted a lot of research attention. Many recent works have focused on proposing new architectures to improve ViT and deploying it into real-world applications. However, little effort has been made to analyze and understand ViT’s architecture design space and its implication of hardware-cost on different devices. In this work, by simply scaling ViT's depth, width, input size, and other basic configurations, we show that a scaled vanilla ViT model without bells and whistles can achieve comparable or superior accuracy-efficiency trade-off than most of the latest ViT variants. Specifically, compared to

DeiT-Tiny, our scaled model achieves a $\uparrow1.9\%$ higher ImageNet top-1 accuracy under the same FLOPs and a $\uparrow3.7\%$ better ImageNet top-1 accuracy under the same latency on an NVIDIA Edge GPU TX2. Motivated by this, we further investigate the extracted scaling strategies from the following two aspects: (1) "can these scaling strategies be transferred across different real hardware devices?''; and (2) "can these scaling strategies be transferred to different ViT variants and tasks?''. For (1), our exploration, based on various devices with different resource budgets, indicates that the transferability effectiveness depends on the underlying device together with its corresponding deployment tool; for (2), we validate the effective transferability of the aforementioned scaling strategies obtained from a vanilla ViT model on top of an image classification task to the PiT model, a strong ViT variant targeting efficiency, as well as object detection and video classification tasks. In particular, when transferred to PiT, our scaling strategies lead to a boosted ImageNet top-1 accuracy of from 74.6% to 76.7% ($\uparrow2.1\%$) under the same 0.7G FLOPs; and when transferred to the COCO object detection task, the average precision is boosted by $\uparrow0.7\%$ under a similar throughput on a V100 GPU.

****

RMNet: Equivalently Removing Residual Connection from Networks

Fanxu Meng,Hao Cheng,Jia-Xin Zhuang,Ke Li,Xing Sun

Although residual connection enables training very deep neural networks, it is not friendly for online inference due to its multi-branch topology. This encourages many researchers to work on designing DNNs without residual connections at inference. For example, RepVGG re-parameterizes multi-branch topology to a VGG-like (single-branch) model when deploying, showing great performance when the network is relatively shallow. However, RepVGG can not transform ResNet to VGG equivalently because re-parameterizing methods can only be applied to linear Blocks and the non-linear layers (ReLU) have to be put outside of the residual connection which results in limited representation ability, especially for deeper networks. In this paper, we aim to remedy this problem and propose to remove the residual connection in a vanilla ResNet equivalently by a reserving and merging (RM) operation on ResBlock. Specifically, RM operation allows input feature maps to pass through the block while reserving their information and merges all the information at the end of each block, which can remove residual connection without changing original output. RMNet basically has two advantages: 1) it achieves a better accuracy-speed trade-off compared with ResNet and RepVGG; 2) its implementation makes it naturally friendly for high ratio network pruning. Extensive experiments are performed to verify the effectiveness of RMNet. We believe the ideology of RMNet can inspire many insights on model design for the community in the future.

****

Few-Shot Attribute Learning

Mengye Ren,Eleni Triantafillou,Kuan-Chieh Wang,James Lucas,Jake Snell,Xaq Pitkow,Andreas S. Tolias,Richard Zemel

Semantic concepts are often defined by a combination of attributes. The use of attributes also facilitates learning of new concepts with zero or few examples. However, the zero-shot learning paradigm assumes that the set of attributes are known and fixed, which is a limitation if a test-time task depends on a previously irrelevant attribute. In this work we study rapid learning of attributes that are previously not labeled in the dataset. Compared to standard few-shot learning of semantic classes, learning new attributes imposes a stiffer challenge. We found that directly supervising the model with a set of training attributes does not generalize well on the test attributes, whereas self-supervised pre-training brings significant improvement. We further experimented with random splits of the attribute space and found that the predictability of attributes provides an informative estimate of a model's ability to generalize.

****

Online Facility Location with Predictions

Shaofeng H.-C. Jiang,Erzhi Liu,You Lyu,Zhihao Gavin Tang,Yubo Zhang

We provide nearly optimal algorithms for online facility location (OFL) with pre dictions. In OFL, n demand points arrive in order and the algorithm must irrevocably assign each demand point to an open facility upon its arrival. The objective is to minimize the total connection costs from demand points to assigned facilities plus the facility opening cost. We further assume the algorithm is additionally given for each demand point x_i a natural prediction $f_{x_i}^{\mathrm{pred}}$ which is supposed to be the facility $f_{x_i}^{\mathrm{opt}}$ that serves x_i in the offline optimal solution.

Our main result is an $O(\min\{\log {\frac{n\eta_\infty}{\mathrm{OPT}}}, \log{n} \})$-competitive algorithm where η_∞ is the maximum prediction error (i.e., the distance between $f_{x_i}^{\mathrm{pred}}$ and $f_{x_i}^{\mathrm{opt}}$). Our algorithm overcomes the fundamental $\Omega(\frac{\log n}{\log \log n})$ lower bound of OFL (without predictions) when η_∞ is small, and it still maintains $O(\log n)$ ratio even when η_∞ is unbounded. Furthermore, our theoretical analysis is supported by empirical evaluations for the tradeoffs between η_∞ and the competitive ratio on various real datasets of different types.

****

Few-Shot Backdoor Attacks on Visual Object Tracking

Yiming Li,Haoxiang Zhong,Xingjun Ma,Yong Jiang,Shu-Tao Xia

Visual object tracking (VOT) has been widely adopted in mission-critical applications, such as autonomous driving and intelligent surveillance systems. In current practice, third-party resources such as datasets, backbone networks, and training platforms are frequently used to train high-performance VOT models. Whilst these resources bring certain convenience, they also introduce new security threats into VOT models. In this paper, we reveal such a threat where an adversary can easily implant hidden backdoors into VOT models by tempering with the training process. Specifically, we propose a simple yet effective few-shot backdoor attack (FSBA) that optimizes two losses alternately: 1) a \emph{feature loss} defined in the hidden feature space, and 2) the standard \emph{tracking loss}. We show that, once the backdoor is embedded into the target model by our FSBA, it can trick the model to lose track of specific objects even when the \emph{trigger} only appears in one or a few frames. We examine our attack in both digital and physical-world settings and show that it can significantly degrade the performance of state-of-the-art VOT trackers. We also show that our attack is resistant to potential defenses, highlighting the vulnerability of VOT models to potential backdoor attacks.

****

Online Unsupervised Learning of Visual Representations and Categories

Mengye Ren,Tyler R. Scott,Michael Louis Iuzzolino,Michael Curtis Mozer,Richard Zemel

Real world learning scenarios involve a nonstationary distribution of classes with sequential dependencies among the samples, in contrast to the standard machine learning formulation of drawing samples independently from a fixed, typically uniform distribution. Furthermore, real world interactions demand learning on-the-fly from few or no class labels. In this work, we propose an unsupervised model that simultaneously performs online visual representation learning and few-shot learning of new categories without relying on any class labels. Our model is a prototype-based memory network with a control component that determines when to form a new class prototype. We formulate it as an online Gaussian mixture model, where components are created online with only a single new example, and assignments do not have to be balanced, which permits an approximation to natural imbalanced distributions from uncurated raw data. Learning includes a contrastive loss that encourages different views of the same image to be assigned to the same prototype. The result is a mechanism that forms categorical representations of objects in nonstationary environments. Experiments show that our method can learn from an online stream of visual input data and is significantly better at category recognition compared to state-of-the-art self-supervised learning methods.

****

Coresets for Kernel Clustering

Shaofeng H.-C. Jiang,Robert Krauthgamer,Jianing Lou,Yubo Zhang

We devise the first coreset for kernel k-Means, and use it to obtain new, more efficient, algorithms. Kernel k-Means has superior clustering capability compared to classical k-Means particularly when clusters are separable non-linearly, but it also introduces significant computational challenges. We address this computational issue by constructing a coreset, which is a reduced dataset that accurately preserves the clustering costs.

Our main result is the first coreset for kernel k-Means, whose size is independent of the number of input points n, and moreover is constructed in time near-linear in n. This result immediately implies new algorithms for kernel k-Means, such as a $(1+\epsilon)$-approximation in time near-linear in n, and a streaming algorithm using space and update time $\mathrm{poly}(k \epsilon^{-1} \log n)$.

We validate our coreset on various datasets with different kernels. Our coreset performs consistently well, achieving small errors while using very few points. We show that our coresets can speed up kernel k-Means++ (the kernelized version of the widely used k-Means++ algorithm), and we further use this faster kernel k-Means++ for spectral clustering. In both applications, we achieve up to 1000x speedup while the error is comparable to baselines that do not use coresets.

****

Physical Gradients for Deep Learning

Philipp Holl,Nils Thuerey,Vladlen Koltun

Solving inverse problems, such as parameter estimation and optimal control, is a vital part of science. Many experiments repeatedly collect data and employ 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 since the magnitude and direction of the gradients can vary strongly. We propose a novel hybrid training approach that combines higher-order optimization methods with machine learning techniques. We replace the gradient of the physical process by a new construct, referred to as the physical gradient. This also allows us to introduce domain knowledge into training by incorporating priors about the solution space into the gradients. We demonstrate the capabilities of our method on a variety of canonical physical systems, showing that physical gradients yield significant improvements on a wide range of optimization and learning problems.

****

How Do Vision Transformers Work?

Namuk Park,Songkuk Kim

The success of multi-head self-attentions (MSAs) for computer vision is now indisputable. However, little is known about how MSAs work. We present fundamental explanations to help better understand the nature of MSAs. In particular, we demonstrate the following properties of MSAs and Vision Transformers (ViTs): (1) MSAs improve not only accuracy but also generalization by flattening the loss landscapes. Such improvement is primarily attributable to their data specificity, not long-range dependency. On the other hand, ViTs suffer from non-convex losses. Large datasets and loss landscape smoothing methods alleviate this problem; (2) MSAs and Convs exhibit opposite behaviors. For example, MSAs are low-pass filters, but Convs are high-pass filters. Therefore, MSAs and Convs are complementary; (3) Multi-stage neural networks behave like a series connection of small individual models. In addition, MSAs at the end of a stage play a key role in prediction. Based on these insights, we propose AlterNet, a model in which Conv blocks at the end of a stage are replaced with MSA blocks. AlterNet outperforms CNNs not only in large data regimes but also in small data regimes. The code is available at https://github.com/xxxnell/how-do-vits-work.

****

Demystifying How Self-Supervised Features Improve Training from Noisy Labels

Hao Cheng,Zhaowei Zhu,Xing Sun,Yang Liu

The advancement of self-supervised learning (SSL) motivates researchers to apply SSL on other tasks such as learning with noisy labels. Recent literature indicates that methods built on SSL features can substantially improve the performance of learning with noisy labels. Nonetheless, the deeper reasons why (and how) SSL features benefit the training from noisy labels are less understood. In this paper, we study why and how self-supervised features help networks resist label noise using both theoretical analyses and numerical experiments. Our result shows that, given a quality encoder pre-trained from SSL, a simple linear layer trained by the cross-entropy loss is theoretically robust to symmetric label noise. Further, we provide insights for how knowledge distilled from SSL features can alleviate the over-fitting problem. We hope our work provides a better understanding for learning with noisy labels from the perspective of self-supervised learning and can potentially serve as a guideline for further research.

****

Backdoor Defense via Decoupling the Training Process

Kunzhe Huang,Yiming Li,Baoyuan Wu,Zhan Qin,Kui Ren

Recent studies have revealed that deep neural networks (DNNs) are vulnerable to backdoor attacks, where attackers embed hidden backdoors in the DNN model by poisoning a few training samples. The attacked model behaves normally on benign samples, whereas its prediction will be maliciously changed when the backdoor is activated. We reveal that poisoned samples tend to cluster together in the feature space of the attacked DNN model, which is mostly due to the end-to-end supervised training paradigm. Inspired by this observation, we propose a novel backdoor defense via decoupling the original end-to-end training process into three stages. Specifically, we first learn the backbone of a DNN model via \emph{self-supervised learning} based on training samples without their labels. The learned backbone will map samples with the same ground-truth label to similar locations in the feature space. Then, we freeze the parameters of the learned backbone and train the remaining fully connected layers via standard training with all (labeled) training samples. Lastly, to further alleviate side-effects of poisoned samples in the second stage, we remove labels of some `low-credible' samples determined based on the learned model and conduct a \emph{semi-supervised fine-tuning} of the whole model. Extensive experiments on multiple benchmark datasets and DNN models verify that the proposed defense is effective in reducing backdoor threats while preserving high accuracy in predicting benign samples. Our code is available at \url{https://github.com/SCLBD/DBD}.

****

Chaining Data - A Novel Paradigm in Artificial Intelligence Exemplified with NMF based Clustering

Norman J Mapes,Sumeet Dua

In the era of artificial intelligence there is an acceleration of high quality inference from the fusion of data and we have overcome the linking challenge associated with higher order features. We have fundamentally linked together tables of databases for clustering algorithms and expect this paradigm and those related to it to produce many new insights. We propose linked view clustering that is an extension of multi-view clustering by adding complementary and consensus information across linked views of each datapoint. While there are many methods, we focus on non-negative matrix factorization combined with the fusion of linking data in a manner that corresponds to extracting knowledge from the multiple tables of a relational database. It is commonplace to identify hashtag communities on social media by word usage, however there exists troves of data not included but could be. We can incorporate locations by hashtag to improve community detection, this is multiNMF or multiview clustering, but we extend this method to beyond the first link. A general artificial intelligence method to incorporate any table that can be chained backwards has not been done before to our knowledge. We call this linked view NMF or chained view clustering and give the algorithms to perform multiplicative updates and the general solution that can be solved using automatic differentiation such as JAX. We demonstrate how the equations can be interpreted on synthetic data as well as how information flows through the links and as a proof of concept on real data we incorporate word vectors using the me thod on an authorship clustering dataset.

****

Variational methods for simulation-based inference

Manuel Glöckler,Michael Deistler,Jakob H. Macke

We present Sequential Neural Variational Inference (SNVI), an approach to perform Bayesian inference in models with intractable likelihoods. SNVI combines likelihood-estimation (or likelihood-ratio-estimation) with variational inference to achieve a scalable simulation-based inference approach. SNVI maintains the flexibility of likelihood(-ratio) estimation to allow arbitrary proposals for simulations, while simultaneously providing a functional estimate of the posterior distribution without requiring MCMC sampling. We present several variants of SNVI and demonstrate that they are substantially more computationally efficient than previous algorithms, without loss of accuracy on benchmark tasks. We apply SNVI to a neuroscience model of the pyloric network in the crab and demonstrate that it can infer the posterior distribution with one order of magnitude fewer simulations than previously reported. SNVI vastly reduces the computational cost of simulation-based inference while maintaining accuracy and flexibility, making it possible to tackle problems that were previously inaccessible.

****

Transformed CNNs: recasting pre-trained convolutional layers with self-attention

Stéphane d'Ascoli,Levent Sagun,Giulio Biroli,Ari S. Morcos

Vision Transformers (ViT) have recently emerged as a powerful alternative to convolutional networks (CNNs). Although hybrid models attempt to bridge the gap between these two architectures, the self-attention layers they rely on induce a strong computational bottleneck, especially at large spatial resolutions. In this work, we explore the idea of reducing the time spent training these layers by initializing them from pre-trained convolutional layers. This enables us to transition smoothly from any pre-trained CNN to its functionally identical hybrid model, called Transformed CNN (T-CNN). With only 50 epochs of fine-tuning, the resulting T-CNNs demonstrate significant performance gains over the CNN as well as substantially improved robustness. We analyze the representations learnt by theT-CNN, providing deeper insights into the fruitful interplay between convolutions and self-attention.

****

Contextual Fusion For Adversarial Robustness

Aiswarya Akumalla,Seth D Haney,Maxim Bazhenov

Mammalian brains handle complex reasoning tasks in a gestalt manner by integrating information from regions of the brain that are specialized to individual sensory modalities. This allows for improved robustness and better generalization ability. In contrast, deep neural networks are usually designed to process one particular information stream and susceptible to various types of adversarial perturbations. While many methods exist for detecting and defending against adversarial attacks, they do not generalize across a range of attacks and negatively affect performance on clean, unperturbed data. We developed a fusion model using a combination of background and foreground features extracted in parallel from Places-CNN and Imagenet-CNN. We tested the benefits of the fusion approach on preserving adversarial robustness for human perceivable (e.g., Gaussian blur) and network perceivable (e.g., gradient-based) attacks for CIFAR-10 and MS COCO data sets. For gradient based attacks, our results show that fusion allows for significant improvements in classification without decreasing performance on unperturbed data and without need to perform adversarial retraining. Our fused model revealed improvements for Gaussian blur type perturbations as well. The increase in performance from fusion approach depended on the variability of the image contexts; larger increases were seen for classes of images with larger differences in their contexts. We also demonstrate the effect of regularization to bias the classifier decision in the presence of a known adversary. We propose that this biologically inspired approach to integrate information across multiple modalities provides a new way to improve adversarial robustness that can be complementary to current state of the art approaches.

****

Learning to Complete Code with Sketches

Daya Guo,Alexey Svyatkovskiy,Jian Yin,Nan Duan,Marc Brockschmidt,Miltiadis Allamanis

Code completion is usually cast as a language modelling problem, i.e., continuing an input in a left-to-right fashion. However, in practice, some parts of the completion (e.g., string literals) may be very hard to predict, whereas subsequent parts directly follow from the context.

To handle this, we instead consider the scenario of generating code completions with "holes" inserted in places where a model is uncertain. We develop Grammformer, a Transformer-based model that guides the code generation by the programming language grammar, and compare it to a variety of more standard sequence models.

We train the models on code completion for C# and Python given partial code context. To evaluate models, we consider both ROUGE as well as a new metric RegexAcc that measures success of generating completions matching long outputs with as few holes as possible.

In our experiments, Grammformer generates 10-50% more accurate completions compared to traditional generative models and 37-50% longer sketches compared to sketch-generating baselines trained with similar techniques.

****

Reverse Engineering of Imperceptible Adversarial Image Perturbations

Yifan Gong,Yuguang Yao,Yize Li,Yimeng Zhang,Xiaoming Liu,Xue Lin,Sijia Liu

It has been well recognized that neural network based image classifiers are easily fooled by images with tiny perturbations crafted by an adversary. There has been a vast volume of research to generate and defend such adversarial attacks. However, the following problem is left unexplored: How to reverse-engineer adversarial perturbations from an adversarial image? This leads to a new adversarial learning paradigm—Reverse Engineering of Deceptions (RED). If successful, RED allows us to estimate adversarial perturbations and recover the original images. However, carefully crafted, tiny adversarial perturbations are difficult to recover by optimizing a unilateral RED objective. For example, the pure image denoising method may overfit to minimizing the reconstruction error but hardly preserve the classification properties of the true adversarial perturbations. To tackle this challenge, we formalize the RED problem and identify a set of principles crucial to the RED approach design. Particularly, we find that prediction alignment and proper data augmentation (in terms of spatial transformations) are two criteria to achieve a generalizable RED approach. By integrating these RED principles with image denoising, we propose a new Class-Discriminative Denoising based RED framework, termed CDD-RED. Extensive experiments demonstrate the effectiveness of CDD-RED under different evaluation metrics (ranging from the pixel-level, prediction-level to the attribution-level alignment) and a variety of attack generation methods (e.g., FGSM, PGD, CW, AutoAttack, and adaptive attacks).

****

Differentially Private SGD with Sparse Gradients

Junyi Zhu,Matthew B. Blaschko

A large number of recent studies reveal that networks and their optimization updates contain information about potentially private training data. To protect sensitive training data, differential privacy has been adopted in deep learning to provide rigorously defined and measurable privacy. However, differentially private stochastic gradient descent (DP-SGD) requires the injection of an amount of noise that scales with the number of gradient dimensions, while neural networks typically contain millions of parameters. As a result, networks trained with DP-SGD typically have large performance drops compared to non-private training. Recent works propose to first project gradients into a lower dimensional subspace, which is found by application of the power method, and then inject noise in this subspace. Although better performance has been achieved, the use of the power method leads to a significantly increased memory footprint by storing sample gradients, and more computational cost by projection. In this work, we mitigate these disadvantages through a sparse gradient representation. Specifically, we randomly freeze a progressively increasing subset of parameters, which results in spa rse gradient updates while maintaining or increasing accuracy over differentially private baselines. Our experiment shows that we can reduce up to 40\% of the gradient dimension while achieve the same performance within the same training epochs. Additionally, sparsity of the gradient updates is beneficial for decreasing communication overhead when deployed in collaborative training, e.g. federated learning. When we apply our approach across various DP-SGD frameworks, we maintain accuracy while achieve up to 70\% representation sparsity, which proves that our approach is a safe and effective add-on to a variety of methods. We further notice that our approach leads to improvement in accuracy in particular for large networks. Importantly, the additional computational cost of our approach is negligible, and results in reduced computation during training due to lower computational cost in power method iterations.

****

A Geometric Perspective on Variational Autoencoders

Clément Chadebec,Stephanie Allassonniere

In this paper, we propose a geometrical interpretation of the Variational Autoencoder framework. We show that VAEs naturally unveil a Riemannian structure of the learned latent space. Moreover, we show that using these geometrical considerations can significantly improve the generation from the vanilla VAE which can now compete with more advanced VAE models on four benchmark data sets. In particular, we propose a new way to generate samples consisting in sampling from the uniform distribution deriving intrinsically from the Riemannian manifold learned by a VAE. We also stress the proposed method's robustness in the low data regime which is known as very challenging for deep generative models. Finally, we validate the method on a complex neuroimaging data set combining both high dimensional data and low sample sizes.

****

DAB-DETR: Dynamic Anchor Boxes are Better Queries for DETR

Shilong Liu,Feng Li,Hao Zhang,Xiao Yang,Xianbiao Qi,Hang Su,Jun Zhu,Lei Zhang

We present in this paper a novel query formulation using dynamic anchor boxes for DETR (DEtection TRansformer) and offer a deeper understanding of the role of queries in DETR. This new formulation directly uses box coordinates as queries in Transformer decoders and dynamically updates them layer by layer. Using box coordinates not only helps using explicit positional priors to improve the query-to-feature similarity and eliminate the slow training convergence issue in DETR, but also allows us to modulate the positional attention map using the box width and height information. Such a design makes it clear that queries in DETR can be implemented as performing soft ROI pooling layer by layer in a cascade manner. As a result, it leads to the best performance on the MS-COCO benchmark among the DETR-like detection models under the same setting, e.g., AP 45.7\% using ResNet50-DC5 as backbone trained in 50 epochs. We also conducted extensive experiments to confirm our analysis and verify the effectiveness of our methods. Code is available at \url{https://github.com/IDEA-opensource/DAB-DETR}.

****

On the Certified Robustness for Ensemble Models and Beyond

Zhuolin Yang,Linyi Li,Xiaojun Xu,Bhavya Kailkhura,Tao Xie,Bo Li

Recent studies show that deep neural networks (DNN) are vulnerable to adversarial examples, which aim to mislead DNNs by adding perturbations with small magnitude. To defend against such attacks, both empirical and theoretical defense approaches have been extensively studied for a single ML model. In this work, we aim to analyze and provide the certified robustness for ensemble ML models, together with the sufficient and necessary conditions of robustness for different ensemble protocols. Although ensemble models are shown more robust than a single model empirically; surprisingly, we find that in terms of the certified robustness the standard ensemble models only achieve marginal improvement compared to a single model. Thus, to explore the conditions that guarantee to provide certifiably robust ensemble ML models, we first prove that diversified gradient and large confidence margin are sufficient and necessary conditions for certifiably robust ensemble models under the model-smoothness assumption. We then provide the bounded model-smoothness analysis based on the proposed Ensemble-before-Smoothing strat egy. We also prove that an ensemble model can always achieve higher certified robustness than a single base model under mild conditions. Inspired by the theoretical findings, we propose the lightweight Diversity Regularized Training (DRT) to train certifiably robust ensemble ML models. Extensive experiments show that our DRT enhanced ensembles can consistently achieve higher certified robustness than existing single and ensemble ML models, demonstrating the state-of-the-art certified L_2-robustness on MNIST, CIFAR-10, and ImageNet datasets.

****

Efficient Neural Causal Discovery without Acyclicity Constraints

Phillip Lippe,Taco Cohen,Efstratios Gavves

Learning the structure of a causal graphical model using both observational and interventional data is a fundamental problem in many scientific fields. A promising direction is continuous optimization for score-based methods, which, however, require constrained optimization to enforce acyclicity or lack convergence guarantees. In this paper, we present ENCO, an efficient structure learning method for directed, acyclic causal graphs leveraging observational and interventional data. ENCO formulates the graph search as an optimization of independent edge likelihoods, with the edge orientation being modeled as a separate parameter. Consequently, we provide for ENCO convergence guarantees under mild conditions, without having to constrain the score function with respect to acyclicity. In experiments, we show that ENCO can efficiently recover graphs with hundreds of nodes, an order of magnitude larger than what was previously possible, while handling deterministic variables and discovering latent confounders.

****

Equivalent Distance Geometry Error for Molecular Conformation Comparison

Shuwen Yang,Tianyu Wen,Ziyao Li,Guojie Song

\textit{Straight-forward} conformation generation models, which generate 3-D structures directly from input molecular graphs, play an important role in various molecular tasks with machine learning, such as 3D-QSAR and virtual screening in drug design. However, existing loss functions in these models either cost overmuch time or fail to guarantee the equivalence during optimization, which means treating different items unfairly, resulting in poor local geometry in generated conformation. So, we propose \textbf{E}quivalent \textbf{D}istance \textbf{G}eometry \textbf{E}rror (EDGE) to calculate the differential discrepancy between conformations where the essential factors of three kinds in conformation geometry (i.e. bond lengths, bond angles and dihedral angles) are equivalently optimized with certain weights. And in the improved version of our method, the optimization features minimizing linear transformations of atom-pair distances within 3-hop. Extensive experiments show that, compared with existing loss functions, EDGE performs effectively and efficiently in two tasks under the same backbones.

****

First-Order Optimization Inspired from Finite-Time Convergent Flows

Siqi Zhang,Mouhacine Benosman,Orlando Romero,Anoop Cherian

In this paper, we investigate the performance of two first-order optimization algorithms, obtained from forward Euler discretization of finite-time optimization flows. These flows are the rescaled-gradient flow (RGF) and the signed-gradient flow (SGF), and consist of non-Lipscthiz or discontinuous dynamical systems that converge locally in finite time to the minima of gradient-dominated functions. We propose an Euler discretization for these first-order finite-time flows, and provide convergence guarantees, in the deterministic and the stochastic setting. We then apply the proposed algorithms to academic examples, as well as deep neural networks training, where we empirically test their performances on the SVHN dataset. Our results show that our schemes demonstrate faster convergences against standard optimization alternatives.

****

Can standard training with clean images outperform adversarial one in robust accuracy?

Jing Wang,Jiahao Hu,Guanrong Li

The deep learning network has achieved great success in almost every field. Unfortunately, it is very vulnerable to adversarial attacks. A lot of researchers ha ve devoted themselves to making the network robust. The most effective one is adversarial training, where malicious examples are generated and fed to train the network. However, this will incur a big computation load. In this work, we ask: “Can standard training with clean images outperform adversarial one in robust accuracy?” Surprisingly, the answer is YES. This success stems from two innovations. The first is a novel loss function that combines the traditional cross-entropy with the feature smoothing loss that encourages the features in an intermediate layer to be uniform. The collaboration between these terms sets up the grounds for our second innovation, namely Active Defense. When a clean or adversarial image feeds into the network, the defender first adds some random noise, then induces this sample to a new smoother one via promotion of feature smoothing. At that point, it can be classified correctly with high probability. Thus the perturbations carefully generated by the attacker can be diminished. While there is an inevitable clean accuracy drop, it is still comparable with others. The great benefit is the robust accuracy outperforms most of the existing methods and is quite resilient to the increase of perturbation budget. Moreover, adaptive attackers also fail to generate effective adversarial samples as the induced perturbations overweight the initial ones imposed by an adversary.

****

FCause: Flow-based Causal Discovery

Tomas Geffner,Emre Kiciman,Angus Lamb,Martin Kukla,Miltiadis Allamanis,Cheng Zhang

Current causal discovery methods either fail to scale, model only limited forms of functional relationships, or cannot handle missing values. This limits their reliability and applicability. We propose FCause, a new flow-based causal discovery method that addresses these drawbacks. Our method is scalable to both high dimensional as well as large volume of data, is able to model complex nonlinear relationships between variables, and can perform causal discovery under partially observed data. Furthermore, our formulation generalizes existing continuous optimization based causal discovery methods, providing a unified view of such models. We perform an extensive empirical evaluation, and show that FCause achieves state of the art results in several causal discovery benchmarks under different conditions reflecting real-world application needs.

****

Pseudo-Labeled Auto-Curriculum Learning for Semi-Supervised Keypoint Localization

Can Wang,Sheng Jin,Yingda Guan,Wentao Liu,Chen Qian,Ping Luo,Wanli Ouyang

Localizing keypoints of an object is a basic visual problem. However, supervised learning of a keypoint localization network often requires a large amount of data, which is expensive and time-consuming to obtain. To remedy this, there is an ever-growing interest in semi-supervised learning (SSL), which leverages a small set of labeled data along with a large set of unlabeled data. Among these SSL approaches, pseudo-labeling (PL) is one of the most popular. PL approaches apply pseudo-labels to unlabeled data, and then train the model with a combination of the labeled and pseudo-labeled data iteratively. The key to the success of PL is the selection of high-quality pseudo-labeled samples. Previous works mostly select training samples by manually setting a single confidence threshold. We propose to automatically select reliable pseudo-labeled samples with a series of dynamic thresholds, which constitutes a learning curriculum.Extensive experiments on five keypoint localization benchmark datasets demonstrate that the proposed approach significantly outperforms the previous state-of-the-art SSL approaches.

****

Signing the Supermask: Keep, Hide, Invert

Nils Koster,Oliver Grothe,Achim Rettinger

The exponential growth in numbers of parameters of neural networks over the past years has been accompanied by an increase in performance across several fields. However, due to their sheer size, the networks not only became difficult to interpret but also problematic to train and use in real-world applications, since hardware requirements increased accordingly.

Tackling both issues, we present a novel approach that either drops a neural net

work's initial weights or inverts their respective sign.
Put simply, a network is trained by weight selection and inversion without changing their absolute values.
Our contribution extends previous work on masking by additionally sign-inverting the initial weights and follows the findings of the Lottery Ticket Hypothesis.
Through this extension and adaptations of initialization methods, we achieve a pruning rate of up to 99%, while still matching or exceeding the performance of various baseline and previous models.
Our approach has two main advantages.
First, and most notable, signed Supermask models drastically simplify a model's structure, while still performing well on given tasks.
Second, by reducing the neural network to its very foundation, we gain insights into which weights matter for performance.
The code is available on GitHub.

****

Folded Hamiltonian Monte Carlo for Bayesian Generative Adversarial Networks
Narges Pourshahrokhi,Samaneh Kouchaki,Yunpeng Li,Payam M. Barnaghi
Generative Adversarial Networks (GANs) can learn complex distributions over images, audio, and data that are difficult to model. We deploy a Bayesian formulation for unsupervised and semi-supervised GAN learning. We propose Folded Hamiltonian Monte Carlo (F-HMC) within this framework to marginalise the weights of the generators and discriminators. The resulting approach improves the performance by having suitable entropy in generated candidates for generator and discriminators' weights. Our proposed model efficiently approximates the high dimensional data due to its parallel composition, increases the accuracy of generated samples and generates interpretable and diverse candidate samples. We have presented the analytical formulation as well as the mathematical proof of the F-HMC. The performance of our model in terms of autocorrelation of generated samples on converging to a high dimensional multi-modal dataset exhibits the effectiveness of the proposed solution. Experimental results on high-dimensional synthetic multi-modal data and natural image benchmarks, including CIFAR-10, SVHN and ImageNet, show that F-HMC outperforms the state-of-the-art methods in terms of test error rates, runtimes per epoch, inception score and Frechet Inception Distance scores.

****

ACCTS: an Adaptive Model Training Policy for Continuous Classification of Time Series
Chenxi Sun,Moxian Song,Derun Cai,Shenda Hong,Hongyan Li
More and more real-world applications require to classify time series at every time. For example, critical patients should be detected for vital signs and diagnosed at all times to facilitate timely life-saving. For this demand, we propose a new concept, Continuous Classification of Time Series (CCTS), to achieve the high-accuracy classification at every time. Time series always evolves dynamically, changing features introducing the multi-distribution form. Thus, different from the existing one-shot classification, the key of CCTS is to model multiple distributions simultaneously. However, most models are hard to achieve it due to their independent identically distributed premise. If a model learns a new distribution, it will likely forget old ones. And if a model repeatedly learns similar data, it will likely be overfitted. Thus, two main problems are the catastrophic forgetting and the over fitting. In this work, we define CCTS as a continual learning task with the unclear distribution division. But different divisions differently affect two problems and a fixed division rule may become invalid as time series evolves. In order to overcome two main problems and finally achieve CCTS, we propose a novel Adaptive model training policy - ACCTS. Its adaptability represents in two aspects: (1) Adaptive multi-distribution extraction policy. Instead of the fixed rules and the prior knowledge, ACCTS extracts data distributions adaptive to the time series evolution and the model change; (2) Adaptive importance-based replay policy. Instead of reviewing all old distributions, ACCTS only replays the important samples adaptive to the contribution of data to the model. Experiments on four real-world datasets show that our method can classify more accurately than all baselines at every time.

****

Bootstrapping Semantic Segmentation with Regional Contrast

Shikun Liu,Shuaifeng Zhi,Edward Johns,Andrew Davison

We present ReCo, a contrastive learning framework designed at a regional level to assist learning in semantic segmentation. ReCo performs pixel-level contrastive learning on a sparse set of hard negative pixels, with minimal additional memory footprint. ReCo is easy to implement, being built on top of off-the-shelf segmentation networks, and consistently improves performance, achieving more accurate segmentation boundaries and faster convergence. The strongest effect is in semi-supervised learning with very few labels. With ReCo, we achieve high quality semantic segmentation model, requiring only 5 examples of each semantic class.

****

Learning Controllable Elements Oriented Representations for Reinforcement Learning

Qi Yi,Jiaming Guo,Rui Zhang,Shaohui Peng,Xing Hu,Xishan Zhang,Ke Tang,Zidong Du,Qi Guo,Yunji Chen

Deep Reinforcement Learning (deep RL) has been successfully applied to solve various decision-making problems in recent years. However, the observations in many real-world tasks are often high dimensional and include much task-irrelevant information, limiting the applications of RL algorithms. To tackle this problem, we propose LCER, a representation learning method that aims to provide RL algorithms with compact and sufficient descriptions of the original observations. Specifically, LCER trains representations to retain the controllable elements of the environment, which can reflect the action-related environment dynamics and thus are likely to be task-relevant. We demonstrate the strength of LCER on the DMControl Suite, proving that it can achieve state-of-the-art performance. To the best of our knowledge, LCER is the first representation learning algorithm that enables the pixel-based SAC to outperform state-based SAC on the DMControl 100K benchmark, showing that the obtained representations can match the oracle descriptions ($i.e.$ the physical states) of the environment.

****

Generative Principal Component Analysis

Zhaoqiang Liu,Jiulong Liu,Subhroshekhar Ghosh,Jun Han,Jonathan Scarlett

In this paper, we study the problem of principal component analysis with generative modeling assumptions, adopting a general model for the observed matrix that encompasses notable special cases, including spiked matrix recovery and phase retrieval. The key assumption is that the first principal eigenvector lies near the range of an L-Lipschitz continuous generative model with bounded k-dimensional inputs. We propose a quadratic estimator, and show that it enjoys a statistical rate of order $\sqrt{\frac{k\log L}{m}}$, where m is the number of samples. Moreover, we provide a variant of the classic power method, which projects the calculated data onto the range of the generative model during each iteration. We show that under suitable conditions, this method converges exponentially fast to a point achieving the above-mentioned statistical rate. This rate is conjectured in~\citep{aubin2019spiked,cocola2020nonasymptotic} to be the best possible even when we only restrict to the special case of spiked matrix models. We perform experiments on various image datasets for spiked matrix and phase retrieval models, and illustrate performance gains of our method to the classic power method and the truncated power method devised for sparse principal component analysis.

****

Pareto Policy Pool for Model-based Offline Reinforcement Learning

Yijun Yang,Jing Jiang,Tianyi Zhou,Jie Ma,Yuhui Shi

Online reinforcement learning (RL) can suffer from poor exploration, sparse reward, insufficient data, and overhead caused by inefficient interactions between an immature policy and a complicated environment. Model-based offline RL instead trains an environment model using a dataset of pre-collected experiences so online RL methods can learn in an offline manner by solely interacting with the model. However, the uncertainty and accuracy of the environment model can drastically vary across different state-action pairs so the RL agent may achieve high mode l return but perform poorly in the true environment. Unlike previous works that need to carefully tune the trade-off between the model return and uncertainty in a single objective, we study a bi-objective formulation for model-based offline RL that aims at producing a pool of diverse policies on the Pareto front performing different levels of trade-offs, which provides the flexibility to select the best policy for each realistic environment from the pool. Our method, ''Pareto policy pool (P3)'', does not need to tune the trade-off weight but can produce policies allocated at different regions of the Pareto front. For this purpose, we develop an efficient algorithm that solves multiple bi-objective optimization problems with distinct constraints defined by reference vectors targeting diverse regions of the Pareto front. We theoretically prove that our algorithm can converge to the targeted regions. In order to obtain more Pareto optimal policies without linearly increasing the cost, we leverage the achieved policies as initialization to find more Pareto optimal policies in their neighborhoods. On the D4RL benchmark for offline RL, P3 substantially outperforms several recent baseline methods over multiple tasks, especially when the quality of pre-collected experiences is low.

****

Filling the G_ap_s: Multivariate Time Series Imputation by Graph Neural Networks

Andrea Cini,Ivan Marisca,Cesare Alippi

Dealing with missing values and incomplete time series is a labor-intensive, tedious, inevitable task when handling data coming from real-world applications. Effective spatio-temporal representations would allow imputation methods to reconstruct missing temporal data by exploiting information coming from sensors at different locations. However, standard methods fall short in capturing the nonlinear time and space dependencies existing within networks of interconnected sensors and do not take full advantage of the available - and often strong - relational information. Notably, most state-of-the-art imputation methods based on deep learning do not explicitly model relational aspects and, in any case, do not exploit processing frameworks able to adequately represent structured spatio-temporal data. Conversely, graph neural networks have recently surged in popularity as both expressive and scalable tools for processing sequential data with relational inductive biases. In this work, we present the first assessment of graph neural networks in the context of multivariate time series imputation. In particular, we introduce a novel graph neural network architecture, named GRIN, which aims at reconstructing missing data in the different channels of a multivariate time series by learning spatio-temporal representations through message passing. Empirical results show that our model outperforms state-of-the-art methods in the imputation task on relevant real-world benchmarks with mean absolute error improvements often higher than 20%.

****

Sparse MoEs meet Efficient Ensembles

James Urquhart Allingham,Florian Wenzel,Zelda E Mariet,Basil Mustafa,Joan Puigcerver,Neil Houlsby,Ghassen Jerfel,Vincent Fortuin,Balaji Lakshminarayanan,Jasper Snoek,Dustin Tran,Carlos Riquelme Ruiz,Rodolphe Jenatton

Machine learning models based on the aggregated outputs of submodels, either at the activation or prediction levels, lead to strong performance. We study the interplay of two popular classes of such models: ensembles of neural networks and sparse mixture of experts (sparse MoEs). First, we show that these two approaches have complementary features whose combination is beneficial. Then, we present partitioned batch ensembles, an efficient ensemble of sparse MoEs that takes the best of both classes of models. Extensive experiments on fine-tuned vision transformers demonstrate the accuracy, log-likelihood, few-shot learning, robustness, and uncertainty calibration improvements of our approach over several challenging baselines. Partitioned batch ensembles not only scale to models with up to 2.7B parameters, but also provide larger performance gains for larger models.

****

The Power of Contrast for Feature Learning: A Theoretical Analysis

Wenlong Ji,Zhun Deng,Ryumei Nakada,James Zou,Linjun Zhang

Contrastive learning has achieved state-of-the-art performance in various self-s upervised learning tasks and even outperforms its supervised counterpart. Despite its empirical success, the theoretical understanding of why contrastive learning works is still limited. In this paper, (i) we provably show that contrastive learning outperforms autoencoder, a classical unsupervised learning method, on both feature recovery and downstream tasks; (ii) we also illustrate the role of labeled data in supervised contrastive learning. This provides theoretical support for recent findings that contrastive learning with labels improves the performance of learned representations in the in-domain downstream task, but it can harm the performance in transfer learning. We verify our theory with numerical experiments.

****

Aggressive Q-Learning with Ensembles: Achieving Both High Sample Efficiency and High Asymptotic Performance

Yanqiu Wu,Xinyue Chen,Che Wang,Yiming Zhang,Zijian Zhou,Keith W. Ross

Recently, Truncated Quantile Critics (TQC), using distributional representation of critics, was shown to provide state-of-the-art asymptotic training performance on all environments from the MuJoCo continuous control benchmark suite. Also recently, Randomized Ensemble Double Q-Learning (REDQ), using a high update-to-data ratio and target randomization, was shown to achieve high sample efficiency that is competitive with state-of-the-art model-based methods. In this paper, we propose a novel model-free algorithm, Aggressive Q-Learning with Ensembles (AQE), which improves the sample-efficiency performance of REDQ and the asymptotic performance of TQC, thereby providing overall state-of-the-art performance during all stages of training. Moreover, AQE is very simple, requiring neither distributional representation of critics nor target randomization.

****

An Unconstrained Layer-Peeled Perspective on Neural Collapse

Wenlong Ji,Yiping Lu,Yiliang Zhang,Zhun Deng,Weijie J Su

Neural collapse is a highly symmetric geometry of neural networks that emerges during the terminal phase of training, with profound implications on the generalization performance and robustness of the trained networks. To understand how the last-layer features and classifiers exhibit this recently discovered implicit bias, in this paper, we introduce a surrogate model called the unconstrained layer-peeled model (ULPM). We prove that gradient flow on this model converges to critical points of a minimum-norm separation problem exhibiting neural collapse in its global minimizer. Moreover, we show that the ULPM with the cross-entropy loss has a benign global landscape for its loss function, which allows us to prove that all the critical points are strict saddle points except the global minimizers that exhibit the neural collapse phenomenon. Empirically, we show that our results also hold during the training of neural networks in real-world tasks when explicit regularization or weight decay is not used.

****

Contrastive Clustering to Mine Pseudo Parallel Data for Unsupervised Translation

Xuan-Phi Nguyen,Hongyu Gong,Yun Tang,Changhan Wang,Philipp Koehn,Shafiq Joty

Modern unsupervised machine translation systems mostly train their models by generating synthetic parallel training data from large unlabeled monolingual corpora of different languages through various means, such as iterative back-translation. However, there may exist small amount of actual parallel data hidden in the sea of unlabeled data, which has not been exploited. We develop a new fine-tuning objective, called Language-Agnostic Constraint for SwAV loss, or LAgSwAV, which enables a pre-trained model to extract such pseudo-parallel data from the monolingual corpora in a fully unsupervised manner. We then propose an effective strategy to utilize the obtained synthetic data to augment unsupervised machine translation. Our method achieves the state of the art in the WMT'14 English-French, WMT'16 German-English and English-Romanian bilingual unsupervised translation tasks, with 40.2, 36.8, 37.0 BLEU, respectively. We also achieve substantial improvements in the FLoRes low-resource English-Nepali and English-Sinhala unsupervised tasks with 5.3 and 5.4 BLEU, respectively.

****

Sound Source Detection from Raw Waveforms with Multi-Scale Synperiodic Filterbanks

Yuhang He

Accurately estimating sound sources' temporal location, spatial location and semantic identity label from multi-channel sound raw waveforms is crucial for an agent to understand the 3D environment acoustically. Multiple sounds form a complex waveform mixture in time, frequency and space, so accurately detecting them requires a representation that can achieve high resolutions across all these dimensions. Existing methods fail to do so because they either extract hand-engineered features\,(i.e. STFT, LogMel) that require a great deal of parameter tuning work (i.e. filter length, window size), or propose to learn a single filter bank to process sound waveforms in a single-scale that often leads to a limited time-frequency resolution capability. In this paper, we tackle this issue by proposing to learn a group of parameterized synperiodic filter banks. Each synperiodic filter's length and frequency response are inversely related, hence is capable of maintaining a better time-frequency resolution trade-off. By alternating the periodicity term, we can easily obtain a group of synperiodic filter banks, where each bank differs in its temporal length. Convolution of the proposed filterbanks with the raw waveform helps to achieve multi-scale perception in the time domain. Moreover, applying synperiodic filter bank to recursively process a downsampled waveform enables us to also achieve multi-scale perception in the frequency domain. Benefiting from the advantage of the multi-scale perception in both time and frequency domain, our proposed synperiodic filter bank groups learn a data-dependent time-frequency resolution map. Following the learnable synperiodic filter bank group front-end, we add a Transformer-like backbone with two parallel soft-stitched branches to learn semantic identity label and spatial location representation semi-independently. Experiments on both direction of arrival estimation task and the physical location estimation task shows our framework outperforms existing methods by a large margin. Replacing existing methods' front-end with synperiodic filter bank also helps to improve the performance.

****

Semi-supervised learning objectives as log-likelihoods in a generative model of data curation

Stoil Krasimirov Ganev,Laurence Aitchison

We currently do not have an understanding of semi-supervised learning (SSL) objectives such as pseudo-labelling and entropy minimization as log-likelihoods, which precludes the development of e.g. Bayesian SSL. Here, we note that benchmark image datasets such as CIFAR-10 are carefully curated, and we formulate SSL objectives as a log-likelihood in a generative model of data curation that was initially developed to explain the cold-posterior effect (Aitchison 2020). SSL methods, from entropy minimization and pseudo-labelling, to state-of-the-art techniques similar to FixMatch can be understood as lower-bounds on our principled log-likelihood. We are thus able to give a proof-of-principle for Bayesian SSL on toy data. Finally, our theory suggests that SSL is effective in part due to the statistical patterns induced by data curation. This provides an explanation of past results which show SSL performs better on clean datasets without any ``out of distribution'' examples. Confirming these results we find that SSL gave much larger performance improvements on curated than on uncurated data, using matched curated and uncurated datasets based on Galaxy Zoo 2.

****

Do What Nature Did To Us: Evolving Plastic Recurrent Neural Networks For Generalized Tasks

Fan Wang,Hao Tian,Haoyi Xiong,hua wu,Yang Cao,Yu Kang,Haifeng Wang

While artificial neural networks (ANNs) have been widely adopted in machine learning, researchers are increasingly obsessed by the gaps between ANNs and natural neural networks (NNNs). In this paper, we propose a framework named as Evolutionary Plastic Recurrent Neural Networks (EPRNN). Inspired by NNN, EPRNN composes Evolution Strategies, Plasticity Rules, and Recursion-based Learning all in one meta learning framework for generalization to different tasks. More specifically, EPRNN incorporates with nested loops for meta learning --- an outer loop searc hes for optimal initial parameters of the neural network and learning rules; an inner loop adapts to specific tasks. In the inner loop of EPRNN, we effectively attain both long term memory and short term memory by forging plasticity with recursion-based learning mechanisms, both of which are believed to be responsible for the formation of memories in NNNs. The inner-loop setting closely simulate that of NNNs, which neither query from any gradient oracle for optimization nor require the exact forms of learning objectives. To evaluate the performance of EPRNN, we carry out extensive experiments in two groups of tasks: Sequence Predicting, and Wheeled Robot Navigating. The experiment results demonstrate the unique advantage of EPRNN compared to state-of-the-arts based on plasticity and recursion while yielding comparably good performance against deep learning based approaches in the tasks. The experiment results suggest the potential of EPRNN to generalize to variety of tasks and encourage more efforts in plasticity and recursion based learning mechanisms.

****

Mitigating Dataset Bias Using Per-Sample Gradients From A Biased Classifier
Sumyeong Ahn,Se-Young Yun

The performance of deep neural networks (DNNs) primarily depends on the configuration of the training set. Specifically, biased training sets can make the trained model have unintended prejudice, which causes severe errors in the inference. Although several studies have addressed biased training using human supervision, few studies have been conducted without human knowledge because biased information cannot be easily extracted without human involvement. This study proposes a simple method to remove prejudice from a biased model without additional information and reconstruct a balanced training set based on the biased training set. The novel training method consists of three steps: (1) training biased DNNs, (2) measuring the contribution to the prejudicial training and generating balanced data batches to prevent the prejudice, (3) training de-biased DNNs with the balanced data. We test the training method based on various synthetic and real-world biased sets and discuss how gradients can efficiently detect minority samples. The experiment demonstrates that the detection method based on the gradients helps erase prejudice, resulting in improved inference accuracy by up to 19.58\% compared to the other state-of-the-art algorithm.

****

Expressiveness of Neural Networks Having Width Equal or Below the Input Dimension
Hans-Peter Beise,Steve Dias Da Cruz

The understanding about the minimum width of deep neural networks needed to ensure universal approximation for different activation functions has progressively been extended \citep{park2020minimum}. In particular, with respect to approximation on general compact sets in the input space, a network width less than or equal to the input dimension excludes universal approximation. In this work, we focus on network functions of width less than or equal to the latter critical bound. We prove a maximum principle from which we conclude that for all continuous and monotonic activation functions, universal approximation of arbitrary continuous functions is impossible on sets that coincide with the boundary of an open set plus an inner point. Conversely, we prove that in this regime, the exact fit of partially constant functions on disjoint compact sets is still possible for ReLU network functions under some conditions on the mutual location of these components. We also show that with cosine as activation function, a three layer network of width one is sufficient to approximate any function on arbitrary finite sets.

****

Deep Ensemble as a Gaussian Process Posterior
Zhijie Deng,Feng Zhou,Jianfei Chen,Guoqiang Wu,Jun Zhu

Deep Ensemble (DE) is a flexible, feasible, and effective alternative to Bayesian neural networks (BNNs) for uncertainty estimation in deep learning. However, DE is broadly criticized for lacking a proper Bayesian justification. Some attempts try to fix this issue, while they are typically coupled with a regression likelihood or rely on restrictive assumptions. In this work, we propose to define a

Gaussian process (GP) approximate posterior with the ensemble members, based on which we perform variational inference directly in the function space. We further develop a function-space posterior regularization mechanism to properly incorporate prior knowledge. We demonstrate the algorithmic benefits of variational inference in the GP family, and provide strategies to make the training feasible. As a result, our method consumes only marginally added training cost than the standard Deep Ensemble. Empirically, our approach achieves better uncertainty estimation than the existing Deep Ensemble and its variants across diverse scenarios.

****

Multimeasurement Generative Models

Saeed Saremi,Rupesh Kumar Srivastava

We formally map the problem of sampling from an unknown distribution with a density in $\mathbb{R}^d$ to the problem of learning and sampling a smoother density in $\mathbb{R}^{Md}$ obtained by convolution with a fixed factorial kernel: the new density is referred to as M-density and the kernel as multimeasurement noise model (MNM). The M-density in $\mathbb{R}^{Md}$ is smoother than the original density in $\mathbb{R}^d$, easier to learn and sample from, yet for large M the two problems are mathematically equivalent since clean data can be estimated exactly given a multimeasurement noisy observation using the Bayes estimator. To formulate the problem, we derive the Bayes estimator for Poisson and Gaussian MNMs in closed form in terms of the unnormalized M-density. This leads to a simple least-squares objective for learning parametric energy and score functions. We present various parametrization schemes of interest including one in which studying Gaussian M-densities directly leads to multidenoising autoencoders—this is the first theoretical connection made between denoising autoencoders and empirical Bayes in the literature. Samples in $\mathbb{R}^d$ are obtained by walk-jump sampling (Saremi & Hyvarinen, 2019) via underdamped Langevin MCMC (walk) to sample from M-density and the multimeasurement Bayes estimation (jump). We study permutation invariant Gaussian M-densities on MNIST, CIFAR-10, and FFHQ-256 datasets, and demonstrate the effectiveness of this framework for realizing fast-mixing stable Markov chains in high dimensions.

****

NAFS: A Simple yet Tough-to-Beat Baseline for Graph Representation Learning

Wentao Zhang,Zeang Sheng,Mingyu Yang,Yang Li,Yu Shen,Zhi Yang,Zichao Yang,Bin CUI

Recently, graph neural networks (GNNs) have shown prominent performance in graph representation learning by leveraging knowledge from both graph structure and node features. However, most of them have two major limitations. First, GNNs can learn higher-order structural information by stacking more layers but can not deal with large depth due to the over-smoothing issue. Second, it is not easy to apply these methods on large graphs due to the expensive computation cost and high memory usage. In this paper, we present node-adaptive feature smoothing (NAFS), a simple non-parametric method that constructs node representations without parameter learning. NAFS first extracts the features of each node with its neighbors of different hops by feature smoothing, and then adaptively combines the smoothed features. Besides, the constructed node representation can further be enhanced by the ensemble of smoothed features extracted via different smoothing strategies. We conduct experiments on four benchmark datasets on two different application scenarios: node clustering and link prediction. Remarkably, NAFS with feature ensemble outperforms the state-of-the-art GNNs on these tasks and mitigates the aforementioned two limitations of most learning-based GNN counterparts.

****

Information Gain Propagation: a New Way to Graph Active Learning with Soft Labels

Wentao Zhang,Yexin Wang,Zhenbang You,Meng Cao,Ping Huang,Jiulong Shan,Zhi Yang,Bin CUI

Graph Neural Networks (GNNs) have achieved great success in various tasks, but their performance highly relies on a large number of labeled nodes, which typically requires considerable human effort. GNN-based Active Learning (AL) methods ar e proposed to improve the labeling efficiency by selecting the most valuable nodes to label. Existing methods assume an oracle can correctly categorize all the selected nodes and thus just focus on the node selection. However, such an exact labeling task is costly, especially when the categorization is out of the domain of individual expert (oracle). The paper goes further, presenting a soft-label approach to AL on GNNs. Our key innovations are: i) relaxed queries where a domain expert (oracle) only judges the correctness of the predicted labels (a binary question) rather than identifying the exact class (a multi-class question), and ii) new criteria of maximizing information gain propagation for active learner with relaxed queries and soft labels. Empirical studies on public datasets demonstrate that our method significantly outperforms the state-of-the-art GNN-based AL methods in terms of both accuracy and labeling cost.

****

Identity-Disentangled Adversarial Augmentation for Self-supervised Learning
Kaiwen Yang,Tianyi Zhou,Xinmei Tian,Dacheng Tao
Data augmentation is critical to contrastive self-supervised learning, whose goal is to distinguish a sample's augmentations (positives) from other samples (negatives). However, strong augmentations may change the sample-identity of the positives, while weak augmentation produces easy positives/negatives leading to nearly-zero loss and ineffective learning. In this paper, we study a simple adversarial augmentation method that can modify training data to be hard positives/negatives without distorting the key information about their original identities. In particular, we decompose a sample x to be its variational auto-encoder (VAE) reconstruction $G(x)$ plus the residual $R(x)=x-G(x)$, where $R(x)$ retains most identity-distinctive information due to an information-theoretic interpretation of the VAE objective. We then adversarially perturb $G(x)$ in the VAE's bottleneck space and adds it back to the original $R(x)$ as an augmentation, which is therefore sufficiently challenging for contrastive learning and meanwhile preserves the sample identity intact. We apply this ``identity-disentangled adversarial augmentation (IDAA)'' to different self-supervised learning methods. On multiple benchmark datasets, IDAA consistently improves both their efficiency and generalization performance. We further show that IDAA learned on a dataset can be transferred to other datasets.

****

Constructing Orthogonal Convolutions in an Explicit Manner
Tan Yu,Jun Li,YUNFENG CAI,Ping Li
Convolutions with orthogonal input-output Jacobian matrix, i.e., orthogonal convolution, have recently attracted substantial attention. A convolution layer with an orthogonal Jacobian matrix is 1-Lipschitz in the 2-norm, making the output robust to the perturbation in input. Meanwhile, an orthogonal Jacobian matrix preserves the gradient norm in back-propagation, which is critical for stable training deep networks. Nevertheless, existing orthogonal convolutions are burdened by high computational costs for preserving orthogonality.
In this work, we exploit the relation between the singular values of the convolution layer's Jacobian and the structure of the convolution kernel. To achieve orthogonality, we explicitly construct the convolution kernel for enforcing all singular values of the convolution layer's Jacobian to be 1s. After training, the explicitly constructed orthogonal (ECO) convolution is constructed only once, and their weights are stored. Then, in evaluation, we only need to load the stored weights of the trained ECO convolution, and the computational cost of ECO convolution is the same as the standard dilated convolution. It is more efficient than the recent state-of-the-art approach, skew orthogonal convolution (SOC) in evaluation. Experiments on CIFAR-10 and CIFAR-100 demonstrate that the proposed ECO convolution is faster than SOC in evaluation while leading to competitive standard and certified robust accuracies.

****

X-model: Improving Data Efficiency in Deep Learning with A Minimax Model
Ximei Wang,Xinyang Chen,Jianmin Wang,Mingsheng Long
To mitigate the burden of data labeling, we aim at improving data efficiency for both classification and regression setups in deep learning. However, the curren

t focus is on classification problems while rare attention has been paid to deep regression, which usually requires more human effort to labeling. Further, due to the intrinsic difference between categorical and continuous label space, the common intuitions for classification, \textit{e.g.} cluster assumptions or pseudo labeling strategies, cannot be naturally adapted into deep regression. To this end, we first delved into the existing data-efficient methods in deep learning and found that they either encourage invariance to \textit{data stochasticity} (\textit{e.g.}, consistency regularization under different augmentations) or \textit{model stochasticity} (\textit{e.g.}, difference penalty for predictions of models with different dropout). To take the power of both worlds, we propose a novel \Chi-model by simultaneously encouraging the invariance to {data stochasticity} and {model stochasticity}. Further, the \Chi-model plays a minimax game between the feature extractor and task-specific heads to further enhance the invariance to model stochasticity. Extensive experiments verify the superiority of the \Chi-model among various tasks, from a single-value prediction task of age estimation to a dense-value prediction task of keypoint localization, a 2D synthetic and a 3D realistic dataset, as well as a multi-category object recognition task.
****

Certified Robustness for Free in Differentially Private Federated Learning
Chulin Xie,Yunhui Long,Pin-Yu Chen,Krishnaram Kenthapadi,Bo Li
Federated learning (FL) provides an efficient training paradigm to jointly train a global model leveraging data from distributed users.
As the local training data comes from different users who may not be trustworthy, several studies have shown that FL is vulnerable to poisoning attacks where adversaries add malicious data during training. On the other hand, to protect the privacy of users, FL is usually trained in a differentially private way (DPFL). Given these properties of FL, in this paper, we aim to ask: Can we leverage the innate privacy property of DPFL to provide robustness certification against poisoning attacks? Can we further improve the privacy of FL to improve such certification?
To this end, we first investigate both the user-level and instance-level privacy of FL, and propose novel randomization mechanisms and analysis to achieve improved differential privacy.
We then provide two robustness certification criteria: certified prediction and certified attack cost for DPFL on both levels. Theoretically, given different privacy properties of DPFL, we prove their certified robustness under a bounded number of adversarial users or instances.
Empirically, we conduct extensive experiments to verify our theories under different attacks on a range of datasets. We show that the global model with a tighter privacy guarantee always provides stronger robustness certification in terms of the certified attack cost, while may exhibit tradeoffs regarding the certified prediction.
We believe our work will inspire future research of developing certifiably robust DPFL based on its inherent properties.
****

Finetuned Language Models are Zero-Shot Learners
Jason Wei,Maarten Bosma,Vincent Zhao,Kelvin Guu,Adams Wei Yu,Brian Lester,Nan Du,Andrew M. Dai,Quoc V Le
This paper explores a simple method for improving the zero-shot learning abilities of language models. We show that instruction tuning—finetuning language models on a collection of datasets described via instructions—substantially improves zero-shot performance on unseen tasks. We take a 137B parameter pretrained language model and instruction tune it on over 60 NLP datasets verbalized via natural language instruction templates. We evaluate this instruction-tuned model, which we call FLAN, on unseen task types. FLAN substantially improves the performance of its unmodified counterpart and surpasses zero-shot 175B GPT-3 on 20 of 25 datasets that we evaluate. FLAN even outperforms few-shot GPT-3 by a large margin on ANLI, RTE, BoolQ, AI2-ARC, OpenbookQA, and StoryCloze. Ablation studies reveal that number of finetuning datasets, model scale, and natural language instructions are key to the success of instruction tuning.

****

On the Adversarial Robustness of Vision Transformers

Rulin Shao,Zhouxing Shi,Jinfeng Yi,Pin-Yu Chen,Cho-Jui Hsieh

Following the success in advancing natural language processing and understanding, transformers are expected to bring revolutionary changes to computer vision. This work provides the first and comprehensive study on the robustness of vision transformers (ViTs) against adversarial perturbations. Tested on various white-box and transfer attack settings, we find that ViTs possess better adversarial robustness when compared with convolutional neural networks (CNNs). This observation also holds for certified robustness. We summarize the following main observations contributing to the improved robustness of ViTs:

1) Features learned by ViTs contain less low-level information and are more generalizable, which contributes to superior robustness against adversarial perturbations.

2) Introducing convolutional or tokens-to-token blocks for learning low-level features in ViTs can improve classification accuracy but at the cost of adversarial robustness.

3) Increasing the proportion of transformers in the model structure (when the model consists of both transformer and CNN blocks) leads to better robustness. But for a pure transformer model, simply increasing the size or adding layers cannot guarantee a similar effect.

4) Pre-training on larger datasets does not significantly improve adversarial robustness though it is critical for training ViTs.

5) Adversarial training is also applicable to ViT for training robust models.

Furthermore, feature visualization and frequency analysis are conducted for explanation. The results show that ViTs are less sensitive to high-frequency perturbations than CNNs and there is a high correlation between how well the model learns low-level features and its robustness against different frequency-based perturbations.

****

Stein Latent Optimization for Generative Adversarial Networks

Uiwon Hwang,Heeseung Kim,Dahuin Jung,Hyemi Jang,Hyungyu Lee,Sungroh Yoon

Generative adversarial networks (GANs) with clustered latent spaces can perform conditional generation in a completely unsupervised manner. In the real world, the salient attributes of unlabeled data can be imbalanced. However, most of existing unsupervised conditional GANs cannot cluster attributes of these data in their latent spaces properly because they assume uniform distributions of the attributes. To address this problem, we theoretically derive Stein latent optimization that provides reparameterizable gradient estimations of the latent distribution parameters assuming a Gaussian mixture prior in a continuous latent space. Structurally, we introduce an encoder network and novel unsupervised conditional contrastive loss to ensure that data generated from a single mixture component represent a single attribute. We confirm that the proposed method, named Stein Latent Optimization for GANs (SLOGAN), successfully learns balanced or imbalanced attributes and achieves state-of-the-art unsupervised conditional generation performance even in the absence of attribute information (e.g., the imbalance ratio). Moreover, we demonstrate that the attributes to be learned can be manipulated using a small amount of probe data.

****

Contractive error feedback for gradient compression

Bingcong Li,Shuai Zheng,Parameswaran Raman,Anshumali Shrivastava,Georgios B. Giannakis

On-device memory concerns in distributed deep learning are becoming more severe due to i) the growth of model size in multi-GPU training, and ii) the adoption of neural networks for federated learning on IoT devices with limited storage. In such settings, this work deals with memory issues emerging with communication efficient methods. To tackle associated challenges, key advances are that i) instead of EFSGD that inefficiently manages memory, the sweet spot of convergence and memory usage can be attained via what is here termed contractive error feedbac k (ConEF); and, ii) communication efficiency in ConEF should be achieved by biased and allreducable gradient compression. ConEF is validated on various learning tasks that include image classification, language modeling, and machine translation. ConEF saves 80% – 90% of the extra memory in EFSGD with almost no loss on test performance, while also achieving 1.3x – 5x speedup of SGD.

****

Sparse DETR: Efficient End-to-End Object Detection with Learnable Sparsity

Byungseok Roh,JaeWoong Shin,Wuhyun Shin,Saehoon Kim

DETR is the first end-to-end object detector using a transformer encoder-decoder architecture and demonstrates competitive performance but low computational efficiency. The subsequent work, Deformable DETR, enhances the efficiency of DETR by replacing dense attention with deformable attention, which achieves 10x faster convergence and improved performance. Using the multiscale feature to ameliorate performance, however, the number of encoder queries increases by 20x compared to DETR, and the computation cost of the encoder attention remains a bottleneck. We observe that the encoder queries referenced by the decoder account for only 45% of the total, and find out the detection accuracy does not deteriorate significantly even if only the referenced queries are polished in the encoder block. Inspired by this observation, we propose Sparse DETR that selectively updates only the queries expected to be referenced by the decoder, thus help the model effectively detect objects. In addition, we show that applying an auxiliary detection loss on the selected queries in the encoder improves the performance while minimizing computational overhead. We validate that Sparse DETR achieves better performance than Deformable DETR even with only 10% encoder queries on the COCO dataset. Albeit only the encoder queries are sparsified, the total computation cost decreases by 38% and the frames per second (FPS) increases by 42% compared to Deformable DETR. Code will be released.

****

Tackling the Generative Learning Trilemma with Denoising Diffusion GANs

Zhisheng Xiao,Karsten Kreis,Arash Vahdat

A wide variety of deep generative models has been developed in the past decade. Yet, these models often struggle with simultaneously addressing three key requirements including: high sample quality, mode coverage, and fast sampling. We call the challenge imposed by these requirements the generative learning trilemma, as the existing models often trade some of them for others. Particularly, denoising diffusion models have shown impressive sample quality and diversity, but their expensive sampling does not yet allow them to be applied in many real-world applications. In this paper, we argue that slow sampling in these models is fundamentally attributed to the Gaussian assumption in the denoising step which is justified only for small step sizes. To enable denoising with large steps, and hence, to reduce the total number of denoising steps, we propose to model the denoising distribution using a complex multimodal distribution. We introduce denoising diffusion generative adversarial networks (denoising diffusion GANs) that model each denoising step using a multimodal conditional GAN. Through extensive evaluations, we show that denoising diffusion GANs obtain sample quality and diversity competitive with original diffusion models while being 2000$\times$ faster on the CIFAR-10 dataset. Compared to traditional GANs, our model exhibits better mode coverage and sample diversity. To the best of our knowledge, denoising diffusion GAN is the first model that reduces sampling cost in diffusion models to an extent that allows them to be applied to real-world applications inexpensively.

****

DeepDebug: Fixing Python Bugs Using Stack Traces, Backtranslation, and Code Skeletons

Dawn Drain,Colin Clement,Guillermo Serrato Castilla,Neel Sundaresan

The joint task of bug localization and program repair is an integral part of the software development process. In this work we present DeepDebug, an approach to automated debugging using large, pretrained transformers. We begin by training a bug-creation model on reversed commit data for the purpose of generating synthetic bugs. We apply these synthetic bugs toward two ends. First, we directly tra

in a backtranslation model on all functions from 200K repositories. Next, we focus on 10K repositories for which we can execute tests, and create buggy versions of all functions in those repositories that are covered by passing tests. This provides us with rich debugging information such as stack traces and print statements, which we use to finetune our model which was pretrained on raw source code. Finally, we strengthen all our models by expanding the context window beyond the buggy function itself, and adding a skeleton consisting of that function's parent class, imports, signatures, docstrings, and method bodies, in order of priority. On the QuixBugs benchmark, we increase the total number of fixes found by over 50%, while also decreasing the false positive rate from 35% to 5% and decreasing the timeout from six hours to one minute. On our own benchmark of executable tests, our model fixes 68% of all bugs on its first attempt without using traces, and after adding traces it fixes 75% on first attempt.
****

Does Entity Abstraction Help Generative Transformers Reason?
Nicolas Gontier,Siva Reddy,Christopher Pal
Pre-trained language models (LMs) often struggle to reason logically or generalize in a compositional fashion. Recent work suggests that incorporating external entity knowledge can improve language models' abilities to reason and generalize. However the effect of explicitly providing entity abstraction remains unclear, especially with recent studies suggesting that pre-trained models already encode some of that knowledge in their parameters. In this work, we study the utility of incorporating entity type abstractions into pre-trained Transformers and test these methods on three different NLP tasks requiring different forms of logical reasoning: (1) compositional language understanding with text-based relational reasoning (CLUTRR), (2) multi-hop question answering (HotpotQA), and (3) conversational question answering (CoQA). We propose and empirically explore three different ways to add such abstraction: (i) as additional input embeddings, (ii) as a separate sequence to encode, and (iii) as an auxiliary prediction task for the model. Overall our analysis demonstrate that models with abstract entity knowledge performs slightly better than without it. However, our experiments also show that the benefits strongly depend on the technique used and the task at hand. The best abstraction aware model achieved an overall accuracy of 88.8% compared to the baseline model achieving 62.3% on CLUTRR. In addition, abstraction-aware models showed improved compositional generalization in both interpolation and extrapolation settings. However, for HotpotQA and CoQA, we find that F1 scores improve by only 0.5% on average. Our results suggest that the benefits of explicit abstraction could be very significant in formally defined logical reasoning settings such as CLUTRR, but point to the notion that explicit abstraction is likely less beneficial for NLP tasks having less formal logical structure.
****

Online Target Q-learning with Reverse Experience Replay: Efficiently finding the Optimal Policy for Linear MDPs
Naman Agarwal,Syomantak Chaudhuri,Prateek Jain,Dheeraj Mysore Nagaraj,Praneeth Netrapalli
Q-learning is a popular Reinforcement Learning (RL) algorithm which is widely used in practice with function approximation (Mnih et al., 2015). In contrast, existing theoretical results are pessimistic about Q-learning. For example, (Baird, 1995) shows that Q-learning does not converge even with linear function approximation for linear MDPs. Furthermore, even for tabular MDPs with synchronous updates, Q-learning was shown to have sub-optimal sample complexity (Li et al., 2021, Azar et al., 2013). The goal of this work is to bridge the gap between practical success of Q-learning and the relatively pessimistic theoretical results. The starting point of our work is the observation that in practice, Q-learning is used with two important modifications: (i) training with two networks, called online network and target network simultaneously (online target learning, or OTL) , and (ii) experience replay (ER) (Mnih et al., 2015). While they have been observed to play a significant role in the practical success of Q-learning, a thorough theoretical understanding of how these two modifications improve the convergence behavior of Q-learning has been missing in literature. By carefully combining

the Q-learning with OTL and reverse experience replay (RER) (a form of experience replay), we present novel methods Q-Rex and Q-RexDaRe (Q-Rex+data reuse). We show that Q-Rex efficiently finds the optimal policy for linear MDPs and provide non-asymptotic bounds on sample complexity -- the first such result for a Q-learning method with linear MDPs. Furthermore, we demonstrate that Q-RexDaRe in fact achieves near optimal sample complexity in the tabular setting, improving upon the existing results for vanilla Q-learning.

****

Deep Q-Network with Proximal Iteration
Kavosh Asadi,Rasool Fakoor,Omer Gottesman,Michael Littman,Alex Smola
We employ Proximal Iteration for value-function optimization in reinforcement learning. Proximal Iteration is a computationally efficient technique that enables us to bias the optimization procedure towards more desirable solutions. As a concrete application of Proximal Iteration in deep reinforcement learning, we endow the objective function of the Deep Q-Network (DQN) agent with a proximal term to ensure that the online-network component of DQN remains in the vicinity of the target network. The resultant agent, which we call DQN with Proximal Iteration, or DQNPro, exhibits significant improvements over the original DQN on the Atari benchmark. Our results accentuate the power of employing sound optimization techniques for deep reinforcement learning.

****

Efficient Image Representation Learning with Federated Sampled Softmax
Sagar M. Waghmare,Hang Qi,Huizhong Chen,Mikhail Sirotenko,Tomer Meron
Learning image representations on decentralized data can bring many benefits in cases where data cannot be aggregated across data silos. Softmax cross entropy loss is highly effective and commonly used for learning image representations. Using a large number of classes has proven to be particularly beneficial for the descriptive power of such representations in centralized learning. However, doing so on decentralized data with Federated Learning is not straightforward, as the demand on computation and communication increases proportionally to the number of classes. In this work we introduce Federated Sampled Softmax, a novel resource-efficient approach for learning image representation with Federated Learning. Specifically, the FL clients sample a set of negative classes and optimize only the corresponding model parameters with respect to a sampled softmax objective that approximates the global full softmax objective. We analytically examine the loss formulation and empirically show that our method significantly reduces the number of parameters transferred to and optimized by the client devices, while performing on par with the standard full softmax method. This work creates a possibility for efficiently learning image representations on decentralized data with a large number of classes in a privacy preserving way.

****

Differentially Private Fractional Frequency Moments Estimation with Polylogarithmic Space
Lun Wang,Iosif Pinelis,Dawn Song
We prove that $\mathbb{F}_p$ sketch, a well-celebrated streaming algorithm for frequency moments estimation, is differentially private as is when $p\in(0, 1]$. $\mathbb{F}_p$ sketch uses only polylogarithmic space, exponentially better than existing DP baselines and only worse than the optimal non-private baseline by a logarithmic factor. The evaluation shows that $\mathbb{F}_p$ sketch can achieve reasonable accuracy with strong privacy guarantees. The code for evaluation is included in the supplementary material.

****

SAINT: Improved Neural Networks for Tabular Data via Row Attention and Contrastive Pre-Training
Gowthami Somepalli,Avi Schwarzschild,Micah Goldblum,C. Bayan Bruss,Tom Goldstein
Tabular data underpins numerous high-impact applications of machine learning from fraud detection to genomics and healthcare. Classical approaches to solving tabular problems, such as gradient boosting and random forests, are widely used by practitioners. However, recent deep learning methods have achieved a degree o

f performance competitive with popular techniques. We devise a hybrid deep learning approach to solving tabular data problems. Our method, SAINT, performs attention over both rows and columns, and it includes an enhanced embedding method. We also study a new contrastive self-supervised pre-training method for use when labels are scarce. SAINT consistently improves performance over previous deep learning methods, and it even performs competitively with gradient boosting methods, including XGBoost, CatBoost, and LightGBM, on average over 30 benchmark datasets in regression, binary classification, and multi-class classification tasks.

****

Compressing Transformer-Based Sequence to Sequence Models With Pre-trained Autoencoders for Text Summarization

Ala Alam Falaki,Robin Gras

We proposed a technique to reduce the decoder’s number of parameters in a sequence to sequence (seq2seq) architecture for automatic text summarization. This approach uses a pre-trained AutoEncoder (AE) trained on top of a pre-trained encoder to reduce the encoder’s output dimension and allow to significantly reduce the size of the decoder. The ROUGE score is used to measure the effectiveness of this method by comparing four different latent space dimensionality reductions: 96%, 66%, 50%, 44%. A few well-known frozen pre-trained encoders (BART, BERT, and DistilBERT) have been tested, paired with the respective frozen pre-trained AEs to test the reduced dimension latent space’s ability to train a 3-layer transformer decoder. We also repeated the same experiments on a small transformer model that has been trained for text summarization. This study shows an increase of the R-1 score by 5% while reducing the model size by 44% using the DistilBERT encoder, and competitive scores for all the other models associated to important size reduction.

****

On the Global Convergence of Gradient Descent for multi-layer ResNets in the mean-field regime

Zhiyan Ding,Shi Chen,Qin Li,Stephen Wright

Finding the optimal configuration of parameters in ResNet is a nonconvex minimization problem, but first-order methods nevertheless find the global optimum in the overparameterized regime. We study this phenomenon with mean-field analysis, by translating the training process of ResNet to a gradient-flow partial differential equation (PDE) and examining the convergence properties of this limiting process.

The activation function is assumed to be 2-homogeneous or partially 1-homogeneous; the regularized ReLU satisfies the latter condition. We show that if the ResNet is sufficiently large, with depth and width depending algebraically on the accuracy and confidence levels, first-order optimization methods can find global minimizers that fit the training data.

****

How Low Can We Go: Trading Memory for Error in Low-Precision Training

Chengrun Yang,Ziyang Wu,Jerry Chee,Christopher De Sa,Madeleine Udell

Low-precision arithmetic trains deep learning models using less energy, less memory and less time. However, we pay a price for the savings: lower precision may yield larger round-off error and hence larger prediction error. As applications proliferate, users must choose which precision to use to train a new model, and chip manufacturers must decide which precisions to manufacture. We view these precision choices as a hyperparameter tuning problem, and borrow ideas from meta-learning to learn the tradeoff between memory and error. In this paper, we introduce Pareto Estimation to Pick the Perfect Precision (PEPPP). We use matrix factorization to find non-dominated configurations (the Pareto frontier) with a limited number of network evaluations. For any given memory budget, the precision that minimizes error is a point on this frontier. Practitioners can use the frontier to trade memory for error and choose the best precision for their goals.

****

Calibration Regularized Training of Deep Neural Networks using Kernel Density Estimation

Teodora Popordanoska,Raphael Sayer,Matthew B. Blaschko

Calibrated probabilistic classifiers are models whose predicted probabilities can directly be interpreted as uncertainty estimates. This property is particularly important in safety-critical applications such as medical diagnosis or autonomous driving. However, it has been shown recently that deep neural networks are poorly calibrated and tend to output overconfident predictions. As a remedy, we propose a trainable calibration error estimator based on Dirichlet kernel density estimates, which asymptotically converges to the true Lp calibration error. This novel estimator enables us to achieve the strongest notion of multiclass calibration, called canonical calibration, while other common calibration methods only allow for top-label and marginal calibration. The empirical results show that our estimator is competitive with the state-of-the-art, consistently yielding tradeoffs between calibration error and accuracy that are (near) Pareto optimal across a range of network architectures. The computational complexity of our estimator is O(n^2), matching that of the kernel maximum mean discrepancy, used in a previously considered trainable calibration estimator. By contrast, 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 an estimator.

****

Uniform Generalization Bounds for Overparameterized Neural Networks

Sattar Vakili,Michael Bromberg,Jezabel R Garcia,Da-shan Shiu,Alberto Bernacchia

An interesting observation in artificial neural networks is their favorable generalization error despite typically being extremely overparameterized. It is well known that the classical statistical learning methods often result in vacuous generalization errors in the case of overparameterized neural networks. Adopting the recently developed Neural Tangent (NT) kernel theory, we prove uniform generalization bounds for overparameterized neural networks in kernel regimes, when the true data generating model belongs to the reproducing kernel Hilbert space (RKHS) corresponding to the NT kernel. Importantly, our bounds capture the exact error rates depending on the differentiability of the activation functions. In order to establish these bounds, we propose the information gain of the NT kernel as a measure of complexity of the learning problem. Our analysis uses a Mercer decomposition of the NT kernel in the basis of spherical harmonics and the decay rate of the corresponding eigenvalues. As a byproduct of our results, we show the equivalence between the RKHS corresponding to the NT kernel and its counterpart corresponding to the Matérn family of kernels, showing the NT kernels induce a very general class of models. We further discuss the implications of our analysis for some recent results on the regret bounds for reinforcement learning and bandit algorithms, which use overparameterized neural networks.

****

F8Net: Fixed-Point 8-bit Only Multiplication for Network Quantization

Qing Jin,Jian Ren,Richard Zhuang,Sumant Hanumante,Zhengang Li,Zhiyu Chen,Yanzhi Wang,Kaiyuan Yang,Sergey Tulyakov

Neural network quantization is a promising compression technique to reduce memory footprint and save energy consumption, potentially leading to real-time inference. However, there is a performance gap between quantized and full-precision models. To reduce it, existing quantization approaches require high-precision INT32 or full-precision multiplication during inference for scaling or dequantization. This introduces a noticeable cost in terms of memory, speed, and required energy. To tackle these issues, we present F8Net, a novel quantization framework consisting in only fixed-point 8-bit multiplication. To derive our method, we first discuss the advantages of fixed-point multiplication with different formats of fixed-point numbers and study the statistical behavior of the associated fixed-point numbers. Second, based on the statistical and algorithmic analysis, we apply different fixed-point formats for weights and activations of different layers. We introduce a novel algorithm to automatically determine the right format for each layer during training. Third, we analyze a previous quantization algorithm—parameterized clipping activation (PACT)—and reformulate it using fixed-point arithmetic. Finally, we unify the recently proposed method for quantization fine-tuning a nd our fixed-point approach to show the potential of our method. We verify F8Net on ImageNet for MobileNet V1/V2 and ResNet18/50. Our approach achieves comparable and better performance, when compared not only to existing quantization techniques with INT32 multiplication or floating point arithmetic, but also to the full-precision counterparts, achieving state-of-the-art performance.

****

ScaLA: Speeding-Up Fine-tuning of Pre-trained Transformer Networks via Efficient and Scalable Adversarial Perturbation

Minjia Zhang,Niranjan Uma Naresh,Yuxiong He

The size of transformer networks is growing at an unprecedented rate and has increased by three orders of magnitude in recent years, approaching trillion-level parameters. To train models of increasing sizes, researchers and practitioners have employed large-batch optimization to leverage massive distributed deep learning systems and resources. However, increasing the batch size changes the training dynamics, often leading to generalization gap and training instability issues that require extensive hyperparameter turning to maintain the same level of accuracy. In this paper, we explore the steepness of the loss landscape of large-batch optimization and find that it tends to be highly complex and irregular, posing challenges to generalization. To address this challenge, we propose ScaLA, a scalable and robust method for large-batch optimization of transformer networks via adversarial perturbation. In particular, we take a sequential game-theoretic approach to make large-batch optimization robust to adversarial perturbation, which helps smooth the loss landscape and improve generalization. Moreover, we perform several optimizations to reduce the computational cost from adversarial perturbation, improving its performance and scalability in the distributed training environment.

We provide a theoretical convergence rate analysis for ScaLA using techniques for analyzing non-convex saddle-point problems. Finally, we perform an extensive evaluation of our method using BERT and RoBERTa on GLUE datasets. Our results show that our method attains up to 18 $\times$ fine-tuning speedups on 2 DGX-2 nodes, while achieving comparable and sometimes higher accuracy than the state-of-the-art large-batch optimization methods. When using the same number of hardware resources, ScaLA is 2.7--9.8$\times$ faster than the baselines.

****

In a Nutshell, the Human Asked for This: Latent Goals for Following Temporal Specifications

Borja G. León,Murray Shanahan,Francesco Belardinelli

We address the problem of building agents whose goal is to learn to execute out-of distribution (OOD) multi-task instructions expressed in temporal logic (TL) by using deep reinforcement learning (DRL). Recent works provided evidence that the agent's neural architecture is a key feature when DRL agents are learning to solve OOD tasks in TL. Yet, the studies on this topic are still in their infancy. In this work, we propose a new deep learning configuration with inductive biases that lead agents to generate latent representations of their current goal, yielding a stronger generalization performance. We use these latent-goal networks within a neuro-symbolic framework that executes multi-task formally-defined instructions and contrast the performance of the proposed neural networks against employing different state-of-the-art (SOTA) architectures when generalizing to unseen instructions in OOD environments.

****

Transform2Act: Learning a Transform-and-Control Policy for Efficient Agent Design

Ye Yuan,Yuda Song,Zhengyi Luo,Wen Sun,Kris M. Kitani

An agent's functionality is largely determined by its design, i.e., skeletal structure and joint attributes (e.g., length, size, strength). However, finding the optimal agent design for a given function is extremely challenging since the problem is inherently combinatorial and the design space is prohibitively large. Additionally, it can be costly to evaluate each candidate design which requires solving for its optimal controller. To tackle these problems, our key idea is to incorporate the design procedure of an agent into its decision-making process. S pecifically, we learn a conditional policy that, in an episode, first applies a sequence of transform actions to modify an agent's skeletal structure and joint attributes, and then applies control actions under the new design. To handle a variable number of joints across designs, we use a graph-based policy where each graph node represents a joint and uses message passing with its neighbors to output joint-specific actions. Using policy gradient methods, our approach enables joint optimization of agent design and control as well as experience sharing across different designs, which improves sample efficiency substantially. Experiments show that our approach, Transform2Act, outperforms prior methods significantly in terms of convergence speed and final performance. Notably, Transform2Act can automatically discover plausible designs similar to giraffes, squids, and spiders. Code and videos are available at https://sites.google.com/view/transform2act.

****

Discrete Representations Strengthen Vision Transformer Robustness

Chengzhi Mao,Lu Jiang,Mostafa Dehghani,Carl Vondrick,Rahul Sukthankar,Irfan Essa

Vision Transformer (ViT) is emerging as the state-of-the-art architecture for image recognition. While recent studies suggest that ViTs are more robust than their convolutional counterparts, our experiments find that ViTs are overly reliant on local features (\eg, nuisances and texture) and fail to make adequate use of global context (\eg, shape and structure). As a result, ViTs fail to generalize to out-of-distribution, real-world data. To address this deficiency, we present a simple and effective architecture modification to ViT's input layer by adding discrete tokens produced by a vector-quantized encoder. Different from the standard continuous pixel tokens, discrete tokens are invariant under small perturbations and contain less information individually, which promote ViTs to learn global information that is invariant. Experimental results demonstrate that adding discrete representation on four architecture variants strengthens ViT robustness by up to 12\% across seven ImageNet robustness benchmarks while maintaining the performance on ImageNet.

****

Imbedding Deep Neural Networks

Andrew Corbett,Dmitry Kangin

Continuous-depth neural networks, such as Neural ODEs, have refashioned the understanding of residual neural networks in terms of non-linear vector-valued optimal control problems. The common solution is to use the adjoint sensitivity method to replicate a forward-backward pass optimisation problem. We propose a new approach which explicates the network's `depth' as a fundamental variable, thus reducing the problem to a system of forward-facing initial value problems. This new method is based on the principal of `Invariant Imbedding' for which we prove a general solution, applicable to all non-linear, vector-valued optimal control problems with both running and terminal loss.

Our new architectures provide a tangible tool for inspecting the theoretical--and to a great extent unexplained--properties of network depth. They also constitute a resource of discrete implementations of Neural ODEs comparable to classes of imbedded residual neural networks. Through a series of experiments, we show the competitive performance of the proposed architectures for supervised learning and time series prediction.

****

On the Convergence of the Monte Carlo Exploring Starts Algorithm for Reinforcement Learning

Che Wang,Shuhan Yuan,Kai Shao,Keith W. Ross

A simple and natural algorithm for reinforcement learning (RL) is Monte Carlo Exploring Starts (MCES), where the Q-function is estimated by averaging the Monte Carlo returns, and the policy is improved by choosing actions that maximize the current estimate of the Q-function. Exploration is performed by "exploring starts", that is, each episode begins with a randomly chosen state and action, and then follows the current policy to the terminal state. In the classic book on RL by Sutton & Barto (2018), it is stated that establishing convergence for the MCES algorithm is one of the most important remaining open theoretical problems in R

L. However, the convergence question for MCES turns out to be quite nuanced. Bertsekas & Tsitsiklis (1996) provide a counter-example showing that the MCES algorithm does not necessarily converge. Tsitsiklis (2002) further shows that if the original MCES algorithm is modified so that the Q-function estimates are updated at the same rate for all state-action pairs, and the discount factor is strictly less than one, then the MCES algorithm converges.

In this paper we make headway with the original and more efficient MCES algorithm given in Sutton et al. (1998), establishing almost sure convergence for Optimal Policy Feed-Forward MDPs, which are MDPs whose states are not revisited within any episode when using an optimal policy. Such MDPs include a large class of environments such as all deterministic environments and all episodic environments with a timestep or any monotonically changing values as part of the state. Different from the previous proofs using stochastic approximations, we introduce a novel inductive approach, which is very simple and only makes use of the strong law of large numbers.

****

Understanding and Preventing Capacity Loss in Reinforcement Learning

Clare Lyle,Mark Rowland,Will Dabney

The reinforcement learning (RL) problem is rife with sources of non-stationarity that can destabilize or inhibit learning progress.

We identify a key mechanism by which this occurs in agents using neural networks as function approximators: \textit{capacity loss}, whereby networks trained to predict a sequence of target values lose their ability to quickly fit new functions over time.

We demonstrate that capacity loss occurs in a broad range of RL agents and environments, and is particularly damaging to learning progress in sparse-reward tasks. We then present a simple regularizer, Initial Feature Regularization (InFeR), that mitigates this phenomenon by regressing a subspace of features towards its value at initialization, improving performance over a state-of-the-art model-free algorithm in the Atari 2600 suite. Finally, we study how this regularization affects different notions of capacity and evaluate other mechanisms by which it may improve performance.

****

Plan Better Amid Conservatism: Offline Multi-Agent Reinforcement Learning with Actor Rectification

Ling Pan,Longbo Huang,Tengyu Ma,Huazhe Xu

The idea of conservatism has led to significant progress in offline reinforcement learning (RL) where an agent learns from pre-collected datasets. However, it is still an open question to resolve offline RL in the more practical multi-agent setting as many real-world scenarios involve interaction among multiple agents. Given the recent success of transferring online RL algorithms to the multi-agent setting, one may expect that offline RL algorithms will also transfer to multi-agent settings directly. Surprisingly, when conservatism-based algorithms are applied to the multi-agent setting, the performance degrades significantly with an increasing number of agents. Towards mitigating the degradation, we identify that a key issue that the landscape of the value function can be non-concave and policy gradient improvements are prone to local optima. Multiple agents exacerbate the problem since the suboptimal policy by any agent could lead to uncoordinated global failure. Following this intuition, we propose a simple yet effective method, \underline{O}ffline \underline{M}ulti-Agent RL with \underline{A}ctor \underline{R}ectification (OMAR), to tackle this critical challenge via an effective combination of first-order policy gradient and zeroth-order optimization methods for the actor to better optimize the conservative value function. Despite the simplicity, OMAR significantly outperforms strong baselines with state-of-the-art performance in multi-agent continuous control benchmarks.

****

SWARM Parallelism: Training Large Models Can Be Surprisingly Communication-Efficient

Max Ryabinin,Tim Dettmers,Michael Diskin,Alexander Borzunov

Many deep learning applications benefit from using large models with billions of parameters. These models can only be trained with specialized distributed training algorithms that require low-latency and high-bandwidth interconnect. As a result, large models are typically trained in dedicated GPU clusters that can be extremely costly to deploy and operate. In contrast, there are more affordable distributed training setups, such as using cheap "preemptible" instances or pooling together existing resources from multiple regions. However, both these setups come with unique challenges that make it impractical to train large models using conventional model parallelism. In this work, we carefully analyze these challenges and find configurations where training larger models becomes less communication-intensive. Based on these observations, we propose SWARM Parallelism (Stochastically Wired Adaptively Rebalanced Model Parallelism) — a model-parallel training algorithm designed for swarms of poorly connected, heterogeneous unreliable devices. SWARM creates temporary randomized pipelines between available nodes that are rebalanced in case of failure. To further reduce the network usage of our approach, we develop several compression-aware architecture modifications and evaluate their tradeoffs. Finally, we combine our insights to train a large Transformer language model with 1.1B shared parameters (approximately 13B before sharing) on a swarm of preemptible T4 GPUs with less than 400Mb/s network throughput.

****

Isotropic Contextual Representations through Variational Regularization

Cornelia Ferner,Stefan Wegenkittl

Contextual language representations achieve state-of-the-art performance across various natural language processing tasks. However, these representations have been shown to suffer from the degeneration problem, i.e. they occupy a narrow cone in the latent space. This problem can be addressed by enforcing isotropy in the latent space. In analogy to variational autoencoders, we suggest applying a token-level variational loss to a Transformer architecture and introduce the prior distribution's standard deviation as model parameter to optimize isotropy. The encoder-decoder architecture allows for learning interpretable embeddings that can be decoded into text again. Extracted features at sentence-level achieve competitive results on benchmark classification tasks.

****

Interpreting Molecule Generative Models for Interactive Molecule Discovery

Yuanqi Du,Xian Liu,Shengchao Liu,Bolei Zhou

Discovering novel molecules with desired properties is crucial for advancing drug discovery and chemical science. Recently deep generative models can synthesize new molecules by sampling random vectors from latent space and then decoding them to a molecule structure. However, through the feedforward generation pipeline, it is difficult to reveal the underlying connections between latent space and molecular properties as well as customize the output molecule with desired properties. In this work, we develop a simple yet effective method to interpret the latent space of the learned generative models with various molecular properties for more interactive molecule generation and discovery. This method, called Molecular Space Explorer (MolSpacE), is model-agnostic and can work with any pre-trained molecule generative models in an off-the-shelf manner. It first identifies latent directions that govern certain molecular properties via the property separation hyperplane and then moves molecules along the directions for smooth change of molecular structures and properties. This method achieves interactive molecule discovery through identifying interpretable and steerable concepts that emerge in the representations of generative models. Experiments show that MolSpacE can manipulate the output molecule toward desired properties with high success. We further quantify and compare the interpretability of multiple state-of-the-art molecule generative models. An interface and a demo video are developed to illustrate the promising application of interactive molecule discovery.

****

Concurrent Adversarial Learning for Large-Batch Training

Yong Liu,Xiangning Chen,Minhao Cheng,Cho-Jui Hsieh,Yang You

Large-batch training has become a commonly used technique when training neural networks with a large number of GPU/TPU processors. As batch size increases, stoc hastic optimizers tend to converge to sharp local minima, leading to degraded test performance. Current methods usually use extensive data augmentation to increase the batch size, but we found the performance gain with data augmentation decreases as batch size increases, and data augmentation will become insufficient after certain point. In this paper, we propose to use adversarial learning to increase the batch size in large-batch training. Despite being a natural choice for smoothing the decision surface and biasing towards a flat region, adversarial learning has not been successfully applied in large-batch training since it requires at least two sequential gradient computations at each step, which will at least double the running time compared with vanilla training even with a large number of processors. To overcome this issue, we propose a novel Concurrent Adversarial Learning (ConAdv) method that decouple the sequential gradient computations in adversarial learning by utilizing staled parameters. Experimental results demonstrate that ConAdv can successfully increase the batch size on both ResNet-50 and EfficientNet training on ImageNet while maintaining high accuracy. In particular, we show ConAdv along can achieve 75.3\% top-1 accuracy on ImageNet ResNet-50 training with 96K batch size, and the accuracy can be further improved to 76.2\% when combining ConAdv with data augmentation. This is the first work successfully scales ResNet-50 training batch size to 96K.

****

Molecular Graph Representation Learning via Heterogeneous Motif Graph Construction

Zhaoning Yu,Hongyang Gao

We consider feature representation learning of molecular graphs. Graph Neural Networks have been widely used in feature representation learning of molecular graphs. However, most proposed methods focus on the individual molecular graph while neglecting their connections, such as motif-level relationships. We propose a novel molecular graph representation learning method by constructing a Heterogeneous Motif graph (HM-graph) to address this issue. In particular, we build an HM-graph that contains motif nodes and molecular nodes. Each motif node corresponds to a motif extracted from molecules. Then, we propose a Heterogeneous Motif Graph Neural Network (HM-GNN) to learn feature representations for each node in the HM-graph. Our HM-graph also enables effective multi-task learning, especially for small molecular datasets. To address the potential efficiency issue, we propose an edge sampler, which significantly reduces computational resources usage. The experimental results show that our model consistently outperforms previous state-of-the-art models. Under multi-task settings, the promising performances of our methods on combined datasets shed light on a new learning paradigm for small molecular datasets. Finally, we show that our model achieves similar performances with significantly less computational resources by using our edge sampler.

****

Modeling label correlations implicitly through latent label encodings for multi-label text classification

Zhizhong Zeng,Yufen Liu,Wenpeng Gao,Baihong Li,Ting Zhang,Xinguo Yu,Zongkai Yang

Multi-label text classification (MLTC) aims to assign a set of labels to each given document. Unlike single-label text classification methods that often focus on document representation learning, MLTC faces a key challenge of modeling label correlations due to complex label dependencies. Previous state-of-the-art works model label correlations explicitly. It lacks flexibility and is prone to introduce inductive bias that may not always hold, such as label-correlation simplification, sequencing label sets, and label-correlation overload. To address this issue, this paper uses latent label representations to model label correlations implicitly. Specifically, the proposed method concatenates a set of latent labels (instead of actual labels) to the text tokens, inputs them to BERT, then maps the contextual encodings of these latent labels to actual labels cooperatively. The correlations between labels, and between labels and the text are modeled indirectly through these latent-label encodings and their correlations. Such latent and distributed correlation modeling can impose less a priori limits and provide more flexibility. The method is conceptually simple but quite effective. It improves the state-of-the-art results on two widely used benchmark datasets by a l arge margin. Further experiments demonstrate that its effectiveness lies in label-correlation utilization rather than document representation. Feature study reveals the importance of using latent label embeddings. It also reveals that contrary to the other token embeddings, the embeddings of these latent labels are sensitive to tasks; sometimes pretraining them can lead to significant performance loss rather than promotion. This result suggests that they are more related to task information (i.e., the actual labels) than the other tokens.

****

Source-Free Adaptation to Measurement Shift via Bottom-Up Feature Restoration
Cian Eastwood,Ian Mason,Chris Williams,Bernhard Schölkopf

Source-free domain adaptation (SFDA) aims to adapt a model trained on labelled data in a source domain to unlabelled data in a target domain without access to the source-domain data during adaptation. Existing methods for SFDA leverage entropy-minimization techniques which: (i) apply only to classification; (ii) destroy model calibration; and (iii) rely on the source model achieving a good level of feature-space class-separation in the target domain. We address these issues for a particularly pervasive type of domain shift called measurement shift which can be resolved by restoring the source features rather than extracting new ones. In particular, we propose Feature Restoration (FR) wherein we: (i) store a lightweight and flexible approximation of the feature distribution under the source data; and (ii) adapt the feature-extractor such that the approximate feature distribution under the target data realigns with that saved on the source. We additionally propose a bottom-up training scheme which boosts performance, which we call Bottom-Up Feature Restoration (BUFR). On real and synthetic data, we demonstrate that BUFR outperforms existing SFDA methods in terms of accuracy, calibration, and data efficiency, while being less reliant on the performance of the source model in the target domain.

****

ProtoRes: Proto-Residual Network for Pose Authoring via Learned Inverse Kinematics
Boris N. Oreshkin,Florent Bocquelet,Felix G. Harvey,Bay Raitt,Dominic Laflamme

Our work focuses on the development of a learnable neural representation of human pose for advanced AI assisted animation tooling. Specifically, we tackle the problem of constructing a full static human pose based on sparse and variable user inputs (e.g. locations and/or orientations of a subset of body joints). To solve this problem, we propose a novel neural architecture that combines residual connections with prototype encoding of a partially specified pose to create a new complete pose from the learned latent space. We show that our architecture outperforms a baseline based on Transformer, both in terms of accuracy and computational efficiency. Additionally, we develop a user interface to integrate our neural model in Unity, a real-time 3D development platform. Furthermore, we introduce two new datasets representing the static human pose modeling problem, based on high-quality human motion capture data, which will be released publicly along with model code.

****

Multiset-Equivariant Set Prediction with Approximate Implicit Differentiation
Yan Zhang,David W Zhang,Simon Lacoste-Julien,Gertjan J. Burghouts,Cees G. M. Snoek

Most set prediction models in deep learning use set-equivariant operations, but they actually operate on multisets. We show that set-equivariant functions cannot represent certain functions on multisets, so we introduce the more appropriate notion of multiset-equivariance. We identify that the existing Deep Set Prediction Network (DSPN) can be multiset-equivariant without being hindered by set-equivariance and improve it with approximate implicit differentiation, allowing for better optimization while being faster and saving memory. In a range of toy experiments, we show that the perspective of multiset-equivariance is beneficial and that our changes to DSPN achieve better results in most cases. On CLEVR object property prediction, we substantially improve over the state-of-the-art Slot Attention from 8% to 77% in one of the strictest evaluation metrics because of the benefits made possible by implicit differentiation.

****

The Inductive Bias of In-Context Learning: Rethinking Pretraining Example Design

Yoav Levine,Noam Wies,Daniel Jannai,Dan Navon,Yedid Hoshen,Amnon Shashua

Pretraining Neural Language Models (NLMs) over a large corpus involves chunking the text into training examples, which are contiguous text segments of sizes processable by the neural architecture. We highlight a bias introduced by this common practice: we prove that the pretrained NLM can model much stronger dependencies between text segments that appeared in the same training example, than it can between text segments that appeared in different training examples. This intuitive result has a twofold role. First, it formalizes the motivation behind a broad line of recent successful NLM training heuristics, proposed for the pretraining and fine-tuning stages, which do not necessarily appear related at first glance. Second, our result clearly indicates further improvements to be made in NLM pretraining for the benefit of Natural Language Understanding tasks. As an example, we propose ``kNN-Pretraining": we show that including semantically related non-neighboring sentences in the same pretraining example yields improved sentence representations and open domain question answering abilities.■This theoretically motivated degree of freedom for pretraining example design indicates new training schemes for self-improving representations.

****

White Paper Assistance: A Step Forward Beyond the Shortcut Learning

Xuan Cheng,Tianshu Xie,XiaoMin Wang,MingHui Liu,Jiali Deng,Ming Liu

The promising performances of CNNs often overshadow the need to examine whether they are doing in the way we are actually interested. We show through experiments that even over-parameterized models would still solve a dataset by recklessly leveraging spurious correlations, or so-called ``shortcuts’’. To combat with this unintended propensity, we borrow the idea of printer test page and propose a novel approach called White Paper Assistance. Our proposed method is two-fold; (a) we intentionally involves the white paper to detect the extent to which the model has preference for certain characterized patterns and (b) we debias the model by enforcing it to make a random guess on the white paper. We show the consistent accuracy improvements that are manifest in various architectures, datasets and combinations with other techniques. Experiments have also demonstrated the versatility of our approach on imbalanced classification and robustness to corruptions.

****

Learned Simulators for Turbulence

Kim Stachenfeld,Drummond Buschman Fielding,Dmitrii Kochkov,Miles Cranmer,Tobias Pfaff,Jonathan Godwin,Can Cui,Shirley Ho,Peter Battaglia,Alvaro Sanchez-Gonzalez

Turbulence simulation with classical numerical solvers requires high-resolution grids to accurately resolve dynamics. Here we train learned simulators at low spatial and temporal resolutions to capture turbulent dynamics generated at high resolution. We show that our proposed model can simulate turbulent dynamics more accurately than classical numerical solvers at the comparably low resolutions across various scientifically relevant metrics. Our model is trained end-to-end from data and is capable of learning a range of challenging chaotic and turbulent dynamics at low resolution, including trajectories generated by the state-of-the-art Athena++ engine. We show that our simpler, general-purpose architecture outperforms various more specialized, turbulence-specific architectures from the learned turbulence simulation literature. In general, we see that learned simulators yield unstable trajectories; however, we show that tuning training noise and temporal downsampling solves this problem. We also find that while generalization beyond the training distribution is a challenge for learned models, training noise, added loss constraints, and dataset augmentation can help. Broadly, we conclude that our learned simulator outperforms traditional solvers run on coarser grids, and emphasize that simple design choices can offer stability and robust generalization.

****

Modular Lifelong Reinforcement Learning via Neural Composition

Jorge A Mendez,Harm van Seijen,ERIC EATON

Humans commonly solve complex problems by decomposing them into easier subproblems and then combining the subproblem solutions. This type of compositional reasoning permits reuse of the subproblem solutions when tackling future tasks that share part of the underlying compositional structure. In a continual or lifelong reinforcement learning (RL) setting, this ability to decompose knowledge into reusable components would enable agents to quickly learn new RL tasks by leveraging accumulated compositional structures. We explore a particular form of composition based on neural modules and present a set of RL problems that intuitively admit compositional solutions. Empirically, we demonstrate that neural composition indeed captures the underlying structure of this space of problems. We further propose a compositional lifelong RL method that leverages accumulated neural components to accelerate the learning of future tasks while retaining performance on previous tasks via off-line RL over replayed experiences.

****

LMSA: Low-relation Mutil-head Self-Attention Mechanism in Visual Transformer

JingJie Wang,Xiang Wei,Xiaoyu Liu

The Transformer backbone network with the self-attention mechanism as the core has achieved great success in the field of natural language processing and computer vision. However, through the self-attention mechanism brings high performance, it also brings higher computational complexity compared to the classic visual feature extraction methods. To further reduce the complexity of self-attention mechanism and explore its lighter version in computer vision, in this paper, we design a novel lightweighted self-attention mechanism: Low-relation Mutil-head Self-Attention (LMSA), which is superior than the recent self-attention. Specifically, the proposed self-attention mechanism breaks the barrier of the dimensional consistency of the traditional self-attention mechanism, resulting in lower computational complexity and occupies less storage space. In addition, employing the new mechanism can release part of the computing consumption of the Transformer network and make the best use of it. Experimental results show that the dimensional consistency inside the traditional self-attention mechanism is unnecessary. In particular, using Swin as the backbone model for training, the accuracy in CIFAR-10 image classification task is improved by 0.43$\%$, in the meanwhile, the consumption of a single self-attention resource is reduced by 64.58$\%$, and the number of model parameters and model size are reduced by more than 15$\%$. By appropriately compressing the dimensions of the self-attention relationship variables, the Transformer network can be more efficient and even perform better. The results prompt us to rethink the reason why the self-attention mechanism works.

****

Learning Similarity Metrics for Volumetric Simulations with Multiscale CNNs

Georg Kohl,Liwei Chen,Nils Thuerey

Simulations that produce three-dimensional data are ubiquitous in science, ranging from fluid flows to plasma physics. We propose a similarity model based on entropy, which allows for the creation of physically meaningful ground truth distances for the similarity assessment of scalar and vectorial data, produced from transport and motion-based simulations. Utilizing two data acquisition methods derived from this model, we create collections of fields from numerical PDE solvers and existing simulation data repositories, and highlight the importance of an appropriate data distribution for an effective training process. Furthermore, a multiscale CNN architecture that computes a volumetric similarity metric (VolSiM) is proposed. To the best of our knowledge this is the first learning method inherently designed to address the challenges arising for the similarity assessment of high-dimensional simulation data. Additionally, the tradeoff between a large batch size and an accurate correlation computation for correlation-based loss functions is investigated, and the metric's equivariance with respect to rotation and scale operations is analyzed. Finally, the robustness and generalization of VolSiM is evaluated on a large range of test data, as well as a particularly challenging turbulence case study, that is close to potential real-world applications.

****

Maximum Entropy Population Based Training for Zero-Shot Human-AI Coordination

Rui Zhao,Jinming Song,Hu Haifeng,Yang Gao,Yi Wu,Zhongqian Sun,Yang Wei

An AI agent should be able to coordinate with humans to solve tasks. We consider the problem of training a Reinforcement Learning (RL) agent without using any human data, i.e., in a zero-shot setting, to make it capable of collaborating with humans. Standard RL agents learn through self-play. Unfortunately, these agents only know how to collaborate with themselves and normally do not perform well with unseen partners, such as humans. The methodology of how to train a robust agent in a zero-shot fashion is still subject to research. Motivated from the maximum entropy RL, we derive a centralized population entropy objective to facilitate learning of a diverse population of agents, which is later used to train a robust AI agent to collaborate with unseen partners. The proposed method shows its effectiveness compared to baseline methods, including self-play PPO, the standard Population-Based Training (PBT), and trajectory diversity-based PBT, in the popular Overcooked game environment. We also conduct online experiments with real humans and further demonstrate the efficacy of the method in the real world.

****

Optimal ANN-SNN Conversion for High-accuracy and Ultra-low-latency Spiking Neural Networks

Tong Bu,Wei Fang,Jianhao Ding,PENGLIN DAI,Zhaofei Yu,Tiejun Huang

Spiking Neural Networks (SNNs) have gained great attraction due to their distinctive properties of low power consumption and fast inference on neuromorphic hardware. As the most effective method to get deep SNNs, ANN-SNN conversion has achieved comparable performance as ANNs on large-scale datasets. Despite this, it requires long time-steps to match the firing rates of SNNs to the activation of ANNs. As a result, the converted SNN suffers severe performance degradation problems with short time-steps, which hamper the practical application of SNNs. In this paper, we theoretically analyze ANN-SNN conversion error and derive the estimated activation function of SNNs. Then we propose the quantization clip-floor-shift activation function to replace the ReLU activation function in source ANNs, which can better approximate the activation function of SNNs. We prove that the expected conversion error between SNNs and ANNs is zero, enabling us to achieve high-accuracy and ultra-low-latency SNNs. We evaluate our method on CIFAR-10/100 and ImageNet datasets, and show that it outperforms the state-of-the-art ANN-SNN and directly trained SNNs in both accuracy and time-steps. To the best of our knowledge, this is the first time to explore high-performance ANN-SNN conversion with ultra-low latency (4 time-steps). Code is available at https://github.com/putshua/SNN_conversion_QCFS

****

Attend to Who You Are: Supervising Self-Attention for Keypoint Detection and Instance-Aware Association

Sen Yang,Zhicheng Wang,Ze Chen,Yanjie Li,Shoukui Zhang,Zhibin Quan,Shu-Tao Xia,Yiping Bao,Erjin Zhou,Wankou Yang

Bottom-up multi-person pose estimation models need to detect keypoints and learn associative information between keypoints.

We argue that these problems can be entirely solved by the Transformer model. Specifically, the self-attention in Transformer measures the pairwise dependencies between locations, which can play a role in providing association information for keypoints grouping.

However, the naive attention patterns are still not subjectively controlled, so there is no guarantee that the keypoints will always attend to the instances to which they belong.

To address it we propose a novel approach of multi-person keypoint detection and instance association using instance masks to supervise self-attention. By supervising self-attention to be instance-aware, we can assign the detected keypoints to the correct human instances based on the pairwise attention scores, without using pre-defined offset vector fields or embedding like CNN-based bottom-up models. An additional benefit of our method is that the instance segmentation results of any number of people can be directly obtained from the supervised attentio n matrix, thereby simplifying the pixel assignment pipeline.

The experiments on the COCO multi-person keypoint detection challenge and person instance segmentation task demonstrate the effectiveness and simplicity of the proposed method.

****

Using a one dimensional parabolic model of the full-batch loss to estimate learning rates during training

Maximus Mutschler,Kevin Alexander Laube,Andreas Zell

A fundamental challenge in Deep Learning is to find optimal step sizes for stochastic gradient descent automatically. In traditional optimization, line searches are a commonly used method to determine step sizes. One problem in Deep Learning is that finding appropriate step sizes on the full-batch loss is unfeasibly expensive. Therefore, classical line search approaches, designed for losses without inherent noise, are usually not applicable. Recent empirical findings suggest that the full-batch loss behaves locally parabolically in the direction of noisy update step directions. Furthermore, the trend of the optimal update step size changes slowly. By exploiting these findings, this work introduces a line-search method that approximates the full-batch loss with a parabola estimated over several mini-batches. Learning rates are derived from such parabolas during training. In the experiments conducted, our approach mostly outperforms SGD tuned with a piece-wise constant learning rate schedule and other line search approaches for Deep Learning across models, datasets, and batch sizes on validation and test accuracy.

****

Estimating Instance-dependent Label-noise Transition Matrix using DNNs

Shuo Yang,Erkun Yang,Bo Han,Yang Liu,Min Xu,Gang Niu,Tongliang Liu

In label-noise learning, estimating the transition matrix is a hot topic as the matrix plays an important role in building statistically consistent classifiers. Traditionally, the transition from clean labels to noisy labels (i.e., clean label transition matrix) has been widely exploited to learn a clean label classifier by employing the noisy data. Motivated by that classifiers mostly output Bayes optimal labels for prediction, in this paper, we study to directly model the transition from Bayes optimal labels to noisy labels (i.e., Bayes label transition matrix) and learn a classifier to predict Bayes optimal labels. Note that given only noisy data, it is ill-posed to estimate either the clean label transition matrix or the Bayes label transition matrix. But favorably, Bayes optimal labels have less uncertainty compared with the clean labels, i.e., the class posteriors of Bayes optimal labels are one-hot vectors while those of clean labels are not. This enables two advantages to estimate the Bayes label transition matrix, i.e., (a) we could theoretically recover a set of noisy data with Bayes optimal labels under mild conditions; (b) the feasible solution space is much smaller. By exploiting the advantages, we estimate the Bayes label transition matrix by employing a deep neural network in a parameterized way, leading to better generalization and superior classification performance.

****

Emergent Communication at Scale

Rahma Chaabouni,Florian Strub,Florent Altché,Eugene Tarassov,Corentin Tallec,Elnaz Davoodi,Kory Wallace Mathewson,Olivier Tieleman,Angeliki Lazaridou,Bilal Piot

Emergent communication aims for a better understanding of human language evolution and building more efficient representations. We posit that reaching these goals will require scaling up, in contrast to a significant amount of literature that focuses on setting up small-scale problems to tease out desired properties of the emergent languages. We focus on three independent aspects to scale up, namely the dataset, task complexity, and population size. We provide a first set of results for large populations solving complex tasks on realistic large-scale datasets, as well as an easy-to-use codebase to enable further experimentation. In more complex tasks and datasets, we find that RL training can become unstable, but responds well to established stabilization techniques.

We also identify the need for a different metric than topographic similarity, which does not correlate with the generalization performances when working with na tural images. In this context, we probe ease-of-learnability and transfer methods to assess emergent languages. Finally, we observe that larger populations do not induce robust emergent protocols with high generalization performance, leading us to explore different ways to leverage population, through voting and imitation learning.

****

Metrics Matter: A Closer Look on Self-Paced Reinforcement Learning

Pascal Klink,Haoyi Yang,Jan Peters,Joni Pajarinen

Curriculum reinforcement learning (CRL) allows to solve complex tasks by generating a tailored sequence of learning tasks, starting from easy ones and subsequently increasing their difficulty. However, the generation of such task sequences is largely governed by application assumptions, often preventing a theoretical investigation of existing approaches. Recently, Klink et al. (2021) showed how self-paced learning induces a principled interpolation between task distributions in the context of RL, resulting in high learning performance. So far, this interpolation is unfortunately limited to Gaussian distributions. Here, we show that on one side, this parametric restriction is insufficient in many learning cases but that on the other, the interpolation of self-paced RL (SPRL) can be degenerate when not restricted to this parametric form. We show that the introduction of concepts from optimal transport into SPRL prevents aforementioned issues. Experiments demonstrate that the resulting introduction of metric structure into the curriculum allows for a well-behaving non-parametric version of SPRL that leads to stable learning performance across tasks.

****

AS-MLP: An Axial Shifted MLP Architecture for Vision

Dongze Lian,Zehao Yu,Xing Sun,Shenghua Gao

An Axial Shifted MLP architecture (AS-MLP) is proposed in this paper. Different from MLP-Mixer, where the global spatial feature is encoded for information flow through matrix transposition and one token-mixing MLP, we pay more attention to the local features interaction. By axially shifting channels of the feature map, AS-MLP is able to obtain the information flow from different axial directions, which captures the local dependencies. Such an operation enables us to utilize a pure MLP architecture to achieve the same local receptive field as CNN-like architecture. We can also design the receptive field size and dilation of blocks of AS-MLP, \emph{etc}, in the same spirit of convolutional neural networks. With the proposed AS-MLP architecture, our model obtains 83.3\% Top-1 accuracy with 88M parameters and 15.2 GFLOPs on the ImageNet-1K dataset. Such a simple yet effective architecture outperforms all MLP-based architectures and achieves competitive performance compared to the transformer-based architectures (\emph{e.g.}, Swin Transformer) even with slightly lower FLOPs. In addition, AS-MLP is also the first MLP-based architecture to be applied to the downstream tasks (\emph{e.g.}, object detection and semantic segmentation). The experimental results are also impressive. Our proposed AS-MLP obtains 51.5 mAP on the COCO validation set and 49.5 MS mIoU on the ADE20K dataset, which is competitive compared to the transformer-based architectures. Our AS-MLP establishes a strong baseline of MLP-based architecture. Code is available at \url{https://github.com/svip-lab/AS-MLP}.

****

Online Continual Learning on Class Incremental Blurry Task Configuration with Anytime Inference

Hyunseo Koh,Dahyun Kim,Jung-Woo Ha,Jonghyun Choi

Despite rapid advances in continual learning, a large body of research is devoted to improving performance in the existing setups.

While a handful of work do propose new continual learning setups, they still lack practicality in certain aspects.

For better practicality, we first propose a novel continual learning setup that is online, task-free, class-incremental, of blurry task boundaries and subject to inference queries at any moment.

We additionally propose a new metric to better measure the performance of the continual learning methods subject to inference queries at any moment.

To address the challenging setup and evaluation protocol, we propose an effectiv

e method that employs a new memory management scheme and novel learning techniques.
Our empirical validation demonstrates that the proposed method outperforms prior arts by large margins. Code and data splits are available at https://github.com/naver-ai/i-Blurry.
****

Discriminator-Weighted Offline Imitation Learning from Suboptimal Demonstrations
Haoran Xu,Xianyuan Zhan,Honglei Yin,Huiling Qin
We study the problem of offline Imitation Learning (IL) where an agent aims to learn an optimal expert behavior policy without additional online environment interactions. Instead, the agent is provided with a static offline dataset of state-action-next state transition triples from both optimal and non-optimal expert behaviors. This strictly offline imitation learning problem arises in many real-world problems, where environment interactions and expert annotations are costly. Prior works that address the problem either require that expert data occupies the majority proportion of the offline dataset, or need to learn a reward function and perform offline reinforcement learning (RL) based on the learned reward function. In this paper, we propose an imitation learning algorithm to address the problem without additional steps of reward learning and offline RL training for the case when demonstrations containing large-proportion of suboptimal data. Built upon behavioral cloning (BC), we introduce an additional discriminator to distinguish expert and non-expert data, we propose a cooperation strategy to boost the performance of both tasks, this will result in a new policy learning objective and surprisingly, we find its equivalence to a generalized BC objective, where the outputs of discriminator serve as the weights of the BC loss function. Experimental results show that the proposed algorithm can learn behavior policies that are much closer to the optimal policies than policies learned by baseline algorithms.
****

Adversarial Collaborative Learning on Non-IID Features
Qinbin Li,Bingsheng He,Dawn Song
Federated learning has been a popular approach to enable collaborative learning on multiple parties without exchanging raw data. However, the model performance of federated learning may degrade a lot due to non-IID data. While most existing studies focus on non-IID labels, federated learning on non-IID features has largely been overlooked. Different from typical federated learning approaches, the paper proposes a new learning concept called ADCOL (Adversarial Collaborative Learning) for non-IID features. Instead of adopting the widely used model-averaging scheme, ADCOL conducts training in an adversarial way: the server aims to train a discriminator to distinguish the representations of the parties, while the parties aim to generate a common representation distribution. Our experiments on three real-world datasets show that ADCOL achieves better accuracy and is much more communication-efficient than state-of-the-art federated learning algorithms on non-IID features. More importantly, ADCOL points out a promising research direction for collaborative learning.
****

Class-Weighted Evaluation Metrics for Imbalanced Data Classification
Min Du,Nesime Tatbul,Brian Rivers,Akhilesh Kumar Gupta,Lucas Hu,Wei Wang,Ryan Marcus,Shengtian Zhou,Insup Lee,Justin Gottschlich
Class distribution skews in imbalanced datasets may lead to models with prediction bias towards majority classes, making fair assessment of classifiers a challenging task. Metrics such as Balanced Accuracy are commonly used to evaluate a classifier’s prediction performance under such scenarios. However, these metrics fall short when classes vary in importance. In this paper, we propose a simple and general-purpose evaluation framework for imbalanced data classification that is sensitive to arbitrary skews in class cardinalities and importances. Experiments with several state-of-the-art classifiers tested on real-world datasets from three different domains show the effectiveness of our framework – not only in evaluating and ranking classifiers, but also training them.
****

Shift-tolerant Perceptual Similarity Metric

Abhijay Ghildyal,Feng Liu

Existing perceptual similarity metrics assume an image and its reference are well aligned. As a result, these metrics are often sensitive to a small alignment error that is imperceptible to the human eyes. This paper studies the effect of small misalignment, specifically a small shift between the input and reference image, on existing metrics, and accordingly develops a shift-tolerant similarity metric. This paper builds upon LPIPS, a widely used learned perceptual similarity metric and explores architectural design considerations to make it robust against imperceptible misalignment. Specifically, we study a wide spectrum of neural network elements, such as anti-aliasing filtering, pooling, striding, padding, and skip connection, and discuss their roles in making a robust metric. Based on our studies, we develop a new deep neural network-based perceptual similarity metric. Our experiments show that our metric is tolerant to imperceptible shifts while being consistent with the human similarity judgment.

****

Transferring Hierarchical Structure with Dual Meta Imitation Learning

Chongkai Gao,Yizhou Jiang,Feng Chen

Hierarchical Imitation learning (HIL) is an effective way for robots to learn sub-skills from long-horizon unsegmented demonstrations. However, the learned hierarchical structure lacks the mechanism to transfer across multi-tasks or to new tasks, which makes them have to learn from scratch when facing a new situation. Transferring and reorganizing modular sub-skills require fast adaptation ability of the whole hierarchical structure. In this work, we propose Dual Meta Imitation Learning (DMIL), a hierarchical meta imitation learning method where the high-level network and sub-skills are iteratively meta-learned with model-agnostic meta-learning. DMIL uses the likelihood of state-action pairs from each sub-skill as the supervision for the high-level network adaptation, and use the adapted high-level network to determine different data set for each sub-skill adaptation. We theoretically prove the convergence of the iterative training process of DMIL and establish the connection between DMIL and the Expectation-Maximization algorithm. Empirically, we achieve state-of-the-art few-shot imitation learning performance on the meta-world benchmark.

****

Boundary-aware Pre-training for Video Scene Segmentation

Jonghwan Mun,Minchul Shin,Gunsoo Han,Sangho Lee,Seongsu Ha,Joonseok Lee,Eun-Sol Kim

Self-supervised learning has drawn attention through its effectiveness in learning in-domain representations with no ground-truth annotations; in particular, it is shown that properly designed pretext tasks (e.g., contrastive prediction task) bring significant performance gains for a downstream task (e.g., classification task). Inspired from this, we tackle video scene segmentation, which is a task of temporally localizing scene boundaries in a video, with a self-supervised learning framework where we mainly focus on designing effective pretext tasks. In our framework, we discover a pseudo-boundary from a sequence of shots by splitting it into two continuous, non-overlapping sub-sequences and leverage the pseudo-boundary to facilitate the pre-training. Based on this, we introduce three novel boundary-aware pretext tasks: 1) Shot-Scene Matching (SSM), 2) Contextual Group Matching (CGM) and 3) Pseudo-boundary Prediction (PP); SSM and CGM guide the model to maximize intra-scene similarity and inter-scene discrimination while PP encourages the model to identify transitional moments. Through comprehensive analysis, we empirically show that pre-training and transferring contextual representation are both critical to improving the video scene segmentation performance. Lastly, we achieve the new state-of-the-art on the MovieNet-SSeg benchmark. The code will be released.

****

Understanding Generalized Label Smoothing when Learning with Noisy Labels

Jiaheng Wei,Hangyu Liu,Tongliang Liu,Gang Niu,Yang Liu

Label smoothing (LS) is an arising learning paradigm that uses the positively weighted average of both the hard training labels and uniformly distributed soft l abels. It was shown that LS serves as a regularizer for training data with hard labels and therefore improves the generalization of the model. Later it was reported LS even helps with improving robustness when learning with noisy labels. However, we observe that the advantage of LS vanishes when we operate in a high label noise regime. Puzzled by the observation, we proceeded to discover that several proposed learning-with-noisy-labels solutions in the literature instead relate more closely to $\textit{negative label smoothing}$ (NLS), which defines as using a negative weight to combine the hard and soft labels! We show that NLS differs substantially from LS in their achieved model confidence. To differentiate the two cases, we will call LS the positive label smoothing (PLS), and this paper unifies PLS and NLS into $\textit{generalized label smoothing}$ (GLS). We provide understandings for the properties of GLS when learning with noisy labels. Among other established properties, we theoretically show NLS is considered more beneficial when the label noise rates are high. We provide extensive experimental results on multiple benchmarks to support our findings too.

****

Learning with Noisy Labels Revisited: A Study Using Real-World Human Annotations

Jiaheng Wei,Zhaowei Zhu,Hao Cheng,Tongliang Liu,Gang Niu,Yang Liu

Existing research on learning with noisy labels mainly focuses on synthetic label noise. The synthetic noise, though has clean structures which greatly enabled statistical analyses, often fails to model the real-world noise patterns. The recent literature has observed several efforts to offer real-world noisy datasets, e.g., Food-101N, WebVision, and Clothing1M. Yet the existing efforts suffer from two caveats: firstly, the lack of ground-truth verification makes it hard to theoretically study the property and treatment of real-world label noise. Secondly, these efforts are often of large scales, which may result in unfair comparisons of robust methods within reasonable and accessible computation power. To better understand real-world label noise, it is important to establish controllable, easy-to-use, and moderate-sized real-world noisy datasets with both ground-truth and noisy labels. This work presents two new benchmark datasets, which we name as CIFAR-10N, CIFAR-100N (jointly we call them CIFAR-N), equipping the training datasets of CIFAR-10 and CIFAR-100 with human-annotated real-world noisy labels we collected from Amazon Mechanical Turk. We quantitatively and qualitatively show that real-world noisy labels follow an instance-dependent pattern rather than the classically assumed and adopted ones (e.g., class-dependent label noise). We then initiate an effort to benchmarking a subset of the existing solutions using CIFAR-10N and CIFAR-100N. We further proceed to study the memorization of correct and wrong predictions, which further illustrates the difference between human noise and class-dependent synthetic noise. We show indeed the real-world noise patterns impose new and outstanding challenges as compared to synthetic label noise. These observations require us to rethink the treatment of noisy labels, and we hope the availability of these two datasets would facilitate the development and evaluation of future learning with noisy label solutions. The corresponding datasets and the leaderboard are available at http://noisylabels.com.

****

Superclass-Conditional Gaussian Mixture Model For Learning Fine-Grained Embeddings

Jingchao Ni,Wei Cheng,Zhengzhang Chen,Takayoshi Asakura,Tomoya Soma,Sho Kato,Haifeng Chen

Learning fine-grained embeddings is essential for extending the generalizability of models pre-trained on "coarse" labels (e.g., animals). It is crucial to fields for which fine-grained labeling (e.g., breeds of animals) is expensive, but fine-grained prediction is desirable, such as medicine. The dilemma necessitates adaptation of a "coarsely" pre-trained model to new tasks with a few "finer-grained" training labels. However, coarsely supervised pre-training tends to suppress intra-class variation, which is vital for cross-granularity adaptation. In this paper, we develop a training framework underlain by a novel superclass-conditional Gaussian mixture model (SCGM). SCGM imitates the generative process of samples from hierarchies of classes through latent variable modeling of the fine-grained subclasses. The framework is agnostic to the encoders and only adds a few d istribution related parameters, thus is efficient, and flexible to different domains. The model parameters are learned end-to-end by maximum-likelihood estimation via a principled Expectation-Maximization algorithm. Extensive experiments on benchmark datasets and a real-life medical dataset indicate the effectiveness of our method.

****

Optimization inspired Multi-Branch Equilibrium Models

Mingjie Li,Yisen Wang,Xingyu Xie,Zhouchen Lin

Works have shown the strong connections between some implicit models and optimization problems. However, explorations on such relationships are limited. Most works pay attention to some common mathematical properties, such as sparsity. In this work, we propose a new type of implicit model inspired by the designing of the systems' hidden objective functions, called the Multi-branch Optimization induced Equilibrium networks~(MOptEqs). The model architecture is designed based on modelling the hidden objective function for the multi-resolution recognition task. Furthermore, we also propose a new strategy inspired by our understandings of the hidden objective function. In this manner, the proposed model can better utilize the hierarchical patterns for recognition tasks and retain the abilities for interpreting the whole structure as trying to obtain the minima of the problem's goal. Comparing with the state-of-the-art models, our MOptEqs not only enjoys better explainability but are also superior to MDEQ with less parameter consumption and better performance on practical tasks. Furthermore, we also implement various experiments to demonstrate the effectiveness of our new methods and explore the applicability of the model's hidden objective function.

****

Learning to Annotate Part Segmentation with Gradient Matching

Yu Yang,Xiaotian Cheng,Hakan Bilen,Xiangyang Ji

The success of state-of-the-art deep neural networks heavily relies on the presence of large-scale labelled datasets, which are extremely expensive and time-consuming to annotate. This paper focuses on tackling semi-supervised part segmentation tasks by generating high-quality images with a pre-trained GAN and labelling the generated images with an automatic annotator. In particular, we formulate the annotator learning as a learning-to-learn problem. Given a pre-trained GAN, the annotator learns to label object parts in a set of randomly generated images such that a part segmentation model trained on these synthetic images with their predicted labels obtains low segmentation error on a small validation set of manually labelled images. We further reduce this nested-loop optimization problem to a simple gradient matching problem and efficiently solve it with an iterative algorithm. We show that our method can learn annotators from a broad range of labelled images including real images, generated images, and even analytically rendered images. Our method is evaluated with semi-supervised part segmentation tasks and significantly outperforms other semi-supervised competitors when the amount of labelled examples is extremely limited.

****

Vector-quantized Image Modeling with Improved VQGAN

Jiahui Yu,Xin Li,Jing Yu Koh,Han Zhang,Ruoming Pang,James Qin,Alexander Ku,Yuanzhong Xu,Jason Baldridge,Yonghui Wu

Pretraining language models with next-token prediction on massive text corpora has delivered phenomenal zero-shot, few-shot, transfer learning and multi-tasking capabilities on both generative and discriminative language tasks. Motivated by this success, we explore a Vector-quantized Image Modeling (VIM) approach that involves pretraining a Transformer to predict rasterized image tokens autoregressively. The discrete image tokens are encoded from a learned Vision-Transformer-based VQGAN (ViT-VQGAN). We first propose multiple improvements over vanilla VQGAN from architecture to codebook learning, yielding better efficiency and reconstruction fidelity. The improved ViT-VQGAN further improves vector-quantized image modeling tasks, including unconditional, class-conditioned image generation and unsupervised representation learning. When trained on ImageNet at 256x256 resolution, we achieve Inception Score (IS) of 175.1 and Fr'echet Inception Distance (FID) of 4.17, a dramatic improvement over the vanilla VQGAN, which obtains 70.

6 and 17.04 for IS and FID, respectively. Based on ViT-VQGAN and unsupervised pretraining, we further evaluate the pretrained Transformer by averaging intermediate features, similar to Image GPT (iGPT). This ImageNet-pretrained VIM-L significantly beats iGPT-L on linear-probe accuracy from 60.3% to 73.2% for a similar model size. ViM-L also outperforms iGPT-XL which is trained with extra web image data and larger model size.

****

Train Short, Test Long: Attention with Linear Biases Enables Input Length Extrapolation

Ofir Press,Noah Smith,Mike Lewis

Since the introduction of the transformer model by Vaswani et al. (2017), a fundamental question has yet to be answered: how does a model achieve extrapolation at inference time for sequences that are longer than it saw during training? We first show that extrapolation can be enabled by simply changing the position representation method, though we find that current methods do not allow for efficient extrapolation. We therefore introduce a simpler and more efficient position method, Attention with Linear Biases (ALiBi). ALiBi does not add positional embeddings to word embeddings; instead, it biases query-key attention scores with a penalty that is proportional to their distance. We show that this method trains a 1.3 billion parameter model on input sequences of length 1024 that extrapolates to input sequences of length 2048, achieving the same perplexity as a sinusoidal position embedding model trained on inputs of length 2048 but training 11% faster and using 11% less memory. ALiBi's inductive bias towards recency also leads it to outperform multiple strong position methods on the WikiText-103 benchmark.

****

LCS: Learning Compressible Subspaces for Adaptive Network Compression at Inference Time

Maxwell Horton,Elvis Nunez,Anish Prabhu,Anurag Ranjan,Ali Farhadi,Mohammad Rastegari

When deploying deep learning models to a device, it is traditionally assumed that available computational resources (compute, memory, and power) remain static. However, real-world computing systems do not always provide stable resource guarantees. Computational resources need to be conserved when load from other processes is high or battery power is low. Inspired by recent works on neural network subspaces, we propose a method for training a "compressible subspace" of neural networks that contains a fine-grained spectrum of models that range from highly efficient to highly accurate. Our models require no retraining, thus our subspace of models can be deployed entirely on-device to allow adaptive network compression at inference time. We present results for achieving arbitrarily fine-grained accuracy-efficiency trade-offs at inference time for structured and unstructured sparsity. We achieve accuracies on-par with standard models when testing our uncompressed models, and maintain high accuracy for sparsity rates above 90% when testing our compressed models. We also demonstrate that our algorithm extends to quantization at variable bit widths, achieving accuracy on par with individually trained networks.

****

Learning Representation from Neural Fisher Kernel with Low-rank Approximation

Ruixiang ZHANG,Shuangfei Zhai,Etai Littwin,Joshua M. Susskind

In this paper, we study the representation of neural networks from the view of kernels. We first define the Neural Fisher Kernel (NFK), which is the Fisher Kernel applied to neural networks. We show that NFK can be computed for both supervised and unsupervised learning models, which can serve as a unified tool for representation extraction. Furthermore, we show that practical NFKs exhibit low-rank structures. We then propose an efficient algorithm that computes a low-rank approximation of NFK, which scales to large datasets and networks. We show that the low-rank approximation of NFKs derived from unsupervised generative models and supervised learning models gives rise to high-quality compact representations of data, achieving competitive results on a variety of machine learning tasks.

****

Learning Temporally Causal Latent Processes from General Temporal Data
Weiran Yao,Yuewen Sun,Alex Ho,Changyin Sun,Kun Zhang

Our goal is to recover time-delayed latent causal variables and identify their relations from measured temporal data. Estimating causally-related latent variables from observations is particularly challenging as the latent variables are not uniquely recoverable in the most general case. In this work, we consider both a nonparametric, nonstationary setting and a parametric setting for the latent processes and propose two provable conditions under which temporally causal latent processes can be identified from their nonlinear mixtures. We propose LEAP, a theoretically-grounded framework that extends Variational AutoEncoders (VAEs) by enforcing our conditions through proper constraints in causal process prior. Experimental results on various datasets demonstrate that temporally causal latent processes are reliably identified from observed variables under different dependency structures and that our approach considerably outperforms baselines that do not properly leverage history or nonstationarity information. This demonstrates that using temporal information to learn latent processes from their invertible nonlinear mixtures in an unsupervised manner, for which we believe our work is one of the first, seems promising even without sparsity or minimality assumptions.

****

Hyperspherical embedding for novel class classification
Rafael S. Pereira,alexis joly,Patrick Valduriez,Fábio Porto

Deep neural networks proved to be useful to learn representations and perform classification on many different modalities of data. Traditional approaches work well on the closed set problem. For learning tasks involving novel classes, known as the open set problem, the metric learning approach has been proposed. However, while promising, common metric learning approaches require pairwise learning, which significantly increases training cost while adding additional challenges. In this paper we present a method in which the similarity of samples projected onto a feature space is enforced by a metric learning approach without requiring
pairwise evaluation. We compare our approach against known methods in different datasets, achieving results up to 81% more accurate.

****

Maximum Mean Discrepancy for Generalization in the Presence of Distribution and Missingness Shift
Liwen Ouyang,Aaron Key

Covariate shifts are a common problem in predictive modeling on real-world problems. This paper proposes addressing the covariate shift problem by minimizing Maximum Mean Discrepancy (MMD) statistics between the training and test sets in either feature input space, feature representation space, or both. We designed three techniques that we call MMD Representation, MMD Mask, and MMD Hybrid to deal with the scenarios where only a distribution shift exists, only a missingness shift exists, or both types of shift exist, respectively. We find that integrating an MMD loss component helps models use the best features for generalization and avoid dangerous extrapolation as much as possible for each test sample. Models treated with this MMD approach show better performance, calibration, and extrapolation on the test set.

****

Shaped Rewards Bias Emergent Language
Brendon Boldt,Yonatan Bisk,David R Mortensen

One of the primary characteristics of emergent phenomena is that they are determined by the basic properties of the system whence they emerge as opposed to explicitly designed constraints. Reinforcement learning is often used to elicit such phenomena which specifically arise from the pressure to maximize reward. We distinguish two types of rewards. The first is the base reward which is motivated directly by the task being solved. The second is shaped rewards which are designed specifically to make the task easier to learn by introducing biases in the learning process. The inductive bias which reward shaping introduces is problematic for emergent language experimentation because it biases the object of study: th

e emergent language. The fact that shaped rewards are intentionally designed conflicts with the basic premise of emergent phenomena arising from basic principles. In this paper, we use a simple sender-receiver navigation game to demonstrate how reward shaping can 1) explicitly bias the semantics of the learned language, 2) significantly change the entropy of the learned communication, and 3) mask the potential effects of other environmental variables of interest.
****

ProgFed: Effective, Communication, and Computation Efficient Federated Learning by Progressive Training

Hui-Po Wang,Sebastian U Stich,Yang He,Mario Fritz

Federated learning is a powerful distributed learning scheme that allows numerous edge devices to collaboratively train a model without sharing their data. However, training is resource-intensive for edge devices, and limited network bandwidth is often the main bottleneck. Prior work often overcomes the constraints by condensing the models or messages into compact formats, e.g., by gradient compression or distillation. In contrast, we propose ProgFed, the first progressive training framework for efficient and effective federated learning. It inherently reduces computation and two-way communication costs while maintaining the strong performance of the final models. We theoretically prove that ProgFed converges at the same asymptotic rate as standard training on full models. Extensive results on a broad range of architectures, including CNNs (VGG, ResNet, ConvNets) and U-nets, and diverse tasks from simple classification to medical image segmentation show that our highly effective training approach saves up to 20% computation and up to 63% communication costs for converged models. As our approach is also complimentary to prior work on compression, we can achieve a wide range of trade-offs, showing reduced communication of up to $50\times$ at only 0.1% loss in utility.
****

DAdaQuant: Doubly-adaptive quantization for communication-efficient Federated Learning

Robert Hönig,Yiren Zhao,Robert D. Mullins

Federated Learning (FL) is a powerful technique for training a model on a server with data from several clients in a privacy-preserving manner. In FL, a server sends the model to every client, who then train the model locally and send it back to the server. The server aggregates the updated models and repeats the process for several rounds. FL incurs significant communication costs, in particular when transmitting the updated local models from the clients back to the server. Recently proposed algorithms quantize the model parameters to efficiently compress FL communication. These algorithms typically have a quantization level that controls the compression factor. We find that dynamic adaptations of the quantization level can boost compression without sacrificing model quality. First, we introduce a time-adaptive quantization algorithm that increases the quantization level as training progresses. Second, we introduce a client-adaptive quantization algorithm that assigns each individual client the optimal quantization level at every round. Finally, we combine both algorithms into DAdaQuant, the doubly-adaptive quantization algorithm. Our experiments show that DAdaQuant consistently improves client$\rightarrow$server compression, outperforming the strongest non-adaptive baselines by up to $2.8\times$.
****

Diffusion-Based Representation Learning

Korbinian Abstreiter,Stefan Bauer,Bernhard Schölkopf,Arash Mehrjou

Score-based methods represented as stochastic differential equations on a continuous time domain have recently proven successful as a non-adversarial generative model. Training such models relies on denoising score matching, which can be seen as multi-scale denoising autoencoders. Here, we augment the denoising score-matching framework to enable representation learning without any supervised signal. GANs and VAEs learn representations by directly transforming latent codes to data samples. In contrast, the introduced diffusion based representation learning relies on a new formulation of the denoising score-matching objective and thus encodes information needed for denoising. We illustrate how this difference allows for manual control of the level of details encoded in the representation. Using the same approach, we propose to learn an infinite-dimensional latent code which achieves improvements of state-of-the-art models on semi-supervised image classification. As a side contribution, we show how adversarial training in score-based models can improve sample quality and improve sampling speed using a new approximation of the prior at smaller noise scales.

****

Federated Learning with Heterogeneous Architectures using Graph HyperNetworks
Or Litany,Haggai Maron,David Acuna,Jan Kautz,Gal Chechik,Sanja Fidler

Standard Federated Learning (FL) techniques are limited to clients with identical network architectures. As a result, inter-organizational collaboration is severely restricted when both data privacy and architectural proprietary are required. In this work, we propose a new FL framework that removes this limitation by adopting a graph hypernetwork as a shared knowledge aggregator. A property of the graph hyper network is that it can adapt to various computational graphs, thereby allowing meaningful parameter sharing across models. Unlike existing solutions, our framework makes no use of external data and does not require clients to disclose their model architecture. Compared with distillation-based and non-graph hypernetwork baselines, our method performs notably better on standard benchmarks. We additionally show encouraging generalization performance to unseen architectures.

****

The Rich Get Richer: Disparate Impact of Semi-Supervised Learning
Zhaowei Zhu,Tianyi Luo,Yang Liu

Semi-supervised learning (SSL) has demonstrated its potential to improve the model accuracy for a variety of learning tasks when the high-quality supervised data is severely limited. Although it is often established that the average accuracy for the entire population of data is improved, it is unclear how SSL fares with different sub-populations. Understanding the above question has substantial fairness implications when different sub-populations are defined by the demographic groups that we aim to treat fairly. In this paper, we reveal the disparate impacts of deploying SSL: the sub-population who has a higher baseline accuracy without using SSL (the "rich" one) tends to benefit more from SSL; while the sub-population who suffers from a low baseline accuracy (the "poor" one) might even observe a performance drop after adding the SSL module. We theoretically and empirically establish the above observation for a broad family of SSL algorithms, which either explicitly or implicitly use an auxiliary "pseudo-label". Experiments on a set of image and text classification tasks confirm our claims. We introduce a new metric, Benefit Ratio, and promote the evaluation of the fairness of SSL (Equalized Benefit Ratio). We further discuss how the disparate impact can be mitigated. We hope our paper will alarm the potential pitfall of using SSL and encourage a multifaceted evaluation of future SSL algorithms.

****

A Good Representation Detects Noisy Labels
Zhaowei Zhu,Zihao Dong,Hao Cheng,Yang Liu

Label noise is pervasive in real-world datasets, which encodes wrong correlation patterns and impairs the generalization of deep neural networks (DNNs). It is critical to find efficient ways to detect the corrupted patterns. Current methods primarily focus on designing robust training techniques to prevent DNNs from memorizing corrupted patterns. This approach has two outstanding caveats: 1) applying this approach to each individual dataset would often require customized training processes; 2) as long as the model is trained with noisy supervisions, overfitting to corrupted patterns is often hard to avoid, leading to performance drop in detection. In this paper, given good representations, we propose a universally applicable and training-free solution to detect noisy labels. Intuitively, good representations help define "neighbors" of each training instance, and closer instances are more likely to share the same clean label. Based on the neighborhood information, we propose two methods: the first one uses "local voting" via checking the noisy label consensuses of nearby representations. The second one

is a ranking-based approach that scores each instance and filters out a guaranteed number of instances that are likely to be corrupted, again using only representations. Given good (but possibly imperfect) representations that are commonly available in practice, we theoretically analyze how they affect the local voting and provide guidelines for tuning neighborhood size. We also prove the worst-case error bound for the ranking-based method. Experiments with both synthetic and real-world label noise demonstrate our training-free solutions are consistently and significantly improving over most of the training-based baselines.
****

Neural Relational Inference with Node-Specific Information
Ershad Banijamali
Inferring interactions among entities is an important problem in studying dynamical systems, which greatly impacts the performance of downstream tasks, such as prediction. In this paper, we tackle the relational inference problem in a setting where each entity can potentially have a set of individualized information that other entities cannot have access to. Specifically, we represent the system using a graph in which the individualized information become node-specific information (NSI). We build our model in the framework of Neural Relation Inference (NRI), where the interaction among entities are uncovered using variational inference. We adopt NRI model to incorporate the individualized information by introducing private nodes in the graph that represent NSI. Such representation enables us to uncover more accurate relations among the agents and therefore leads to better performance on the downstream tasks. Our experiment results over real-world datasets validate the merit of our proposed algorithm.
****

Representations of Computer Programs in the Human Brain
Shashank Srikant,Benjamin Lipkin,Anna A Ivanova,Evelina Fedorenko,Una-May O'Reilly
We present the first study relating representations of computer programs generated by unsupervised machine learning (ML) models and representations of computer programs in the human brain. We analyze recordings---brain representations---from functional magnetic resonance imaging (fMRI) studies of people comprehending Python code. We discover brain representations, in different and specific regions of the brain, that encode static and dynamic properties of code such as abstract syntax tree (AST)-related information and runtime information. We also map brain representations to representations of a suite of ML models that vary in their complexity. We find that the Multiple Demand system, a system of brain regions previously shown to respond to code, contains information about multiple specific code properties, as well as machine learned representations of code. We make all the corresponding code, data, and analysis publicly available.
****

Deep Fair Discriminative Clustering
Hongjing Zhang,Ian Davidson
Deep clustering has the potential to learn a strong representation and hence better clustering performance than traditional clustering methods such as k-means and spectral clustering. However, this strong representation learning ability may make the clustering unfair by discovering surrogates for protected information which our experiments empirically show. This work studies a general notion of group-level fairness for both binary and multi-state protected status variables (PSVs). We begin by formulating the group-level fairness problem as an integer linear programming whose totally unimodular constraint matrix means it can be efficiently solved via linear programming. We then show how to inject this solver into a discriminative deep clustering backbone and hence propose a refinement learning algorithm to combine the clustering goal with the fairness objective to learn fair clusters adaptively. Experimental results on real-world datasets demonstrate that our model consistently outperforms state-of-the-art fair clustering algorithms. Furthermore, our framework shows promising results for novel fair clustering tasks including flexible fairness constraints, multi-state PSVs, and predictive clustering.
****

Directional Bias Helps Stochastic Gradient Descent to Generalize in Nonparametric Model

Yiling Luo,Xiaoming Huo,Yajun Mei

This paper studies the Stochastic Gradient Descent (SGD) algorithm in kernel regression. The main finding is that SGD with moderate and annealing step size converges in the direction of the eigenvector that corresponds to the largest eigenvalue of the gram matrix. On the contrary, the Gradient Descent (GD) with a moderate or small step size converges along the direction that corresponds to the smallest eigenvalue. For a general squared risk minimization problem, we show that directional bias towards a larger eigenvalue of the Hessian (which is the gram matrix in our case) results in an estimator that is closer to the ground truth. Adopting this result to kernel regression, the directional bias helps the SGD estimator generalize better. This result gives one way to explain how noise helps in generalization when learning with a nontrivial step size, which may be useful for promoting further understanding of stochastic algorithms in deep learning. The correctness of our theory is supported by simulations and experiments of Neural Network on the FashionMNIST dataset.

****

Robust Robotic Control from Pixels using Contrastive Recurrent State-Space Models

Nitish Srivastava,Walter Talbott,Martin Bertran Lopez,Shuangfei Zhai,Joshua M. Susskind

Modeling the world can benefit robot learning by providing a rich training signal for shaping an agent's latent state space. However, learning world models in unconstrained environments over high-dimensional observation spaces such as images is challenging. One source of difficulty is the presence of irrelevant but hard-to-model background distractions, and unimportant visual details of task-relevant entities. We address this issue by learning a recurrent latent dynamics model which contrastively predicts the next observation. This simple model leads to surprisingly robust robotic control even with simultaneous camera, background, and color distractions. We outperform alternatives such as bisimulation methods which impose state-similarity measures derived from divergence in future reward or future optimal actions. We obtain state-of-the-art results on the Distracting Control Suite, a challenging benchmark for pixel-based robotic control.

****

Lower Bounds on the Robustness of Fixed Feature Extractors to Test-time Adversaries

Arjun Nitin Bhagoji,Daniel Cullina,Ben Zhao

Understanding the robustness of machine learning models to adversarial examples generated by test-time adversaries is a problem of great interest. Recent theoretical work has derived lower bounds on how robust \emph{any model} can be, when a data distribution and attacker constraints are specified. However, these bounds only apply to arbitrary classification functions and do not account for specific architectures and models used in practice, such as neural networks. In this paper, we develop a methodology to analyze the robustness of fixed feature extractors, which in turn provide bounds on the robustness of any classifier trained on top of it. In other words, this indicates how robust the representation obtained from that extractor is with respect to a given adversary. Our bounds hold for arbitrary feature extractors. The tightness of these bounds relies on the effectiveness of the method used to find collisions between pairs of perturbed examples at deeper layers. For linear feature extractors, we provide closed-form expressions for collision finding while for arbitrary feature extractors, we propose a bespoke algorithm based on the iterative solution of a convex program that provably finds collisions. We utilize our bounds to identify the layers of robustly trained models that contribute the most to a lack of robustness, as well as compare the same layer across different training methods to provide a quantitative comparison of their relative robustness. Our experiments establish that each of the following lead to a measurable drop in robustness: i) layers that linearly reduce dimension, ii) sparsity induced by ReLU activations and, iii) mismatches in the attacker constraints at train and test time. These findings point towards future design considerations for robust models that arise from our methodology.

****

Bregman Gradient Policy Optimization

Feihu Huang,Shangqian Gao,Heng Huang

In the paper, we design a novel Bregman gradient policy optimization framework for reinforcement learning based on Bregman divergences and momentum techniques. Specifically, we propose a Bregman gradient policy optimization (BGPO) algorithm based on the basic momentum technique and mirror descent iteration. Meanwhile, we further propose an accelerated Bregman gradient policy optimization (VR-BGPO) algorithm based on the variance reduced technique. Moreover, we provide a convergence analysis framework for our Bregman gradient policy optimization under the nonconvex setting. We prove that our BGPO achieves a sample complexity of $O(\epsilon^{-4})$ for finding ϵ-stationary policy only requiring one trajectory at each iteration, and our VR-BGPO reaches the best known sample complexity of $O(\epsilon^{-3})$, which also only requires one trajectory at each iteration. In particular, by using different Bregman divergences, our BGPO framework unifies many existing policy optimization algorithms such as the existing (variance reduced) policy gradient algorithms such as natural policy gradient algorithm. Extensive experimental results on multiple reinforcement learning tasks demonstrate the efficiency of our new algorithms.

****

On Hard Episodes in Meta-Learning

Samyadeep Basu,Amr Sharaf,Nicolo Fusi,Soheil Feizi

Existing meta-learners primarily focus on improving the average task accuracy across multiple episodes. Different episodes, however, may vary in hardness and quality leading to a wide gap in the meta-learner's performance across episodes. Understanding this issue is particularly critical in industrial few-shot settings, where there is limited control over test episodes as they are typically uploaded by end-users. In this paper, we empirically analyse the behaviour of meta-learners on episodes of varying hardness across three standard benchmark datasets: CIFAR-FS, mini-ImageNet, and tiered-ImageNet. Surprisingly, we observe a wide gap in accuracy of around 50% between the hardest and easiest episodes across all the standard benchmarks and meta-learners. We additionally investigate various properties of hard episodes and highlight their connection to catastrophic forgetting during meta-training. To address the issue of sub-par performance on hard episodes, we investigate and benchmark different meta-training strategies based on adversarial training and curriculum learning. We find that adversarial training strategies are much more powerful than curriculum learning in improving the prediction performance on hard episodes.

****

Language Model Pre-training Improves Generalization in Policy Learning

Shuang Li,Xavier Puig,Yilun Du,Ekin Akyürek,Antonio Torralba,Jacob Andreas,Igor Mordatch

Language model (LM) pre-training has proven useful for a wide variety of language processing tasks, including tasks that require nontrivial planning and reasoning capabilities. Can these capabilities be leveraged for more general machine learning problems? We investigate the effectiveness of LM pretraining to scaffold learning and generalization in autonomous decision-making. We use a pre-trained GPT-2 LM to initialize an interactive policy, which we fine-tune via imitation learning to perform interactive tasks in a simulated household environment featuring partial observability, large action spaces, and long time horizons. To leverage pre-training, we first encode observations, goals, and history information as templated English strings, and train the policy to predict the next action. We find that this form of pre-training enables generalization in policy learning: for test tasks involving novel goals or environment states, initializing policies with language models improves task completion rates by nearly 20%. Additional experiments explore the role of language-based encodings in these results; we find that it is possible to train a simple adapter layer that maps from observations and action histories to LM embeddings, and thus that language modeling provides an effective initializer even for tasks with no language as input or output.

Together, these results suggest that language modeling induces representations that are useful for modeling not just language, but natural goals and plans; these representations can aid learning and generalization even outside of language processing.

****

Fully Steerable 3D Spherical Neurons

Pavlo Melnyk,Michael Felsberg,Mårten Wadenbäck

Emerging from low-level vision theory, steerable filters found their counterpart in prior work on steerable convolutional neural networks equivariant to rigid transformations. In our work, we propose a steerable feed-forward learning-based approach that consists of spherical decision surfaces and operates on point clouds. Focusing on 3D geometry, we derive a 3D steerability constraint for hypersphere neurons, which are obtained by conformal embedding of Euclidean space and have recently been revisited in the context of learning representations of point sets. Exploiting the rotational equivariance, we show how our model parameters are fully steerable at inference time. We use a synthetic point set and real-world 3D skeleton data to show how the proposed spherical filter banks enable making equivariant and, after online optimization, invariant class predictions for known point sets in unknown orientations.

****

SPP-RL: State Planning Policy Reinforcement Learning

Jacek Cyranka,Zuzanna Opa■a,Jacek P■ocharczyk,Mikhail Zanka

We introduce an algorithm for reinforcement learning, in which the actor plans for the next state provided the current state. To communicate the actor output to the environment we incorporate an inverse dynamics control model and train it using supervised learning.

We train the RL agent using off-policy state-of-the-art reinforcement learning algorithms: DDPG, TD3, and SAC. To guarantee that the target states are physically relevant, the overall learning procedure is formulated as a constrained optimization problem, solved via the classical Lagrangian optimization method. We benchmark the state planning RL approach using a varied set of continuous environments, including standard MuJoCo tasks, safety-gym level 0 environments, and AntPush. In SPP approach the optimal policy is being searched for in the space of state-state mappings, a considerably larger space than the traditional space of state-action mappings. We report that quite surprisingly SPP implementations attain superior performance to vanilla state-of-the-art off-policy RL algorithms in the tested environments.

****

A generalization of the randomized singular value decomposition

Nicolas Boulle,Alex Townsend

The randomized singular value decomposition (SVD) is a popular and effective algorithm for computing a near-best rank k approximation of a matrix A using matrix-vector products with standard Gaussian vectors. Here, we generalize the theory of randomized SVD to multivariate Gaussian vectors, allowing one to incorporate prior knowledge of A into the algorithm. This enables us to explore the continuous analogue of the randomized SVD for Hilbert--Schmidt (HS) operators using operator-function products with functions drawn from a Gaussian process (GP). We then construct a new covariance kernel for GPs, based on weighted Jacobi polynomials, which allows us to rapidly sample the GP and control the smoothness of the randomly generated functions. Numerical examples on matrices and HS operators demonstrate the applicability of the algorithm.

****

Dropout Q-Functions for Doubly Efficient Reinforcement Learning

Takuya Hiraoka,Takahisa Imagawa,Taisei Hashimoto,Takashi Onishi,Yoshimasa Tsuruoka

Randomized ensembled double Q-learning (REDQ) (Chen et al., 2021b) has recently achieved state-of-the-art sample efficiency on continuous-action reinforcement learning benchmarks. This superior sample efficiency is made possible by using a large Q-function ensemble. However, REDQ is much less computationally efficient than non-ensemble counterparts such as Soft Actor-Critic (SAC) (Haarnoja et al., 2018a). To make REDQ more computationally efficient, we propose a method of improving computational efficiency called DroQ, which is a variant of REDQ that uses a small ensemble of dropout Q-functions. Our dropout Q-functions are simple Q-functions equipped with dropout connection and layer normalization. Despite its simplicity of implementation, our experimental results indicate that DroQ is doubly (sample and computationally) efficient. It achieved comparable sample efficiency with REDQ, much better computational efficiency than REDQ, and comparable computational efficiency with that of SAC.

****

Label Augmentation with Reinforced Labeling for Weak Supervision

Gürkan Solmaz,Flavio Cirillo,Fabio Maresca,Anagha GodeAnilKumar

Weak supervision (WS) is an alternative to the traditional supervised learning to address the need for ground truth. Data programming is a practical WS approach that allows programmatic labeling data samples using labeling functions (LFs) instead of hand-labeling each data point. However, the existing approach fails to fully exploit the domain knowledge encoded into LFs, especially when the LFs' coverage is low. This is due to the common data programming pipeline that neglects to utilize data features during the generative process.

This paper proposes a new approach called reinforced labeling (RL). Given an unlabeled dataset and a set of LFs, RL augments the LFs' outputs to cases not covered by LFs based on similarities among samples. Thus, RL can lead to higher labeling coverage for training an end classifier. The experiments on several domains (classification of YouTube comments, wine quality, and weather prediction) result in considerable gains. The new approach produces significant performance improvement, leading up to +21 points in accuracy and +61 points in F1 scores compared to the state-of-the-art data programming approach.

****

QDrop: Randomly Dropping Quantization for Extremely Low-bit Post-Training Quantization

Xiuying Wei,Ruihao Gong,Yuhang Li,Xianglong Liu,Fengwei Yu

Recently, post-training quantization (PTQ) has driven much attention to produce efficient neural networks without long-time retraining. Despite the low cost, current PTQ works always fail under the extremely low-bit setting. In this study, we pioneeringly confirm that properly incorporating activation quantization into the PTQ reconstruction benefits the final accuracy. To deeply understand the inherent reason, a theoretical framework is established, which inspires us that the flatness of the optimized low-bit model on calibration and test data is crucial. Based on the conclusion, a simple yet effective approach dubbed as \textsc{QDrop} is proposed, which randomly drops the quantization of activations during reconstruction. Extensive experiments on various tasks including computer vision (image classification, object detection) and natural language processing (text classification and question answering) prove its superiority. With \textsc{QDrop}, the limit of PTQ is pushed to the 2-bit activation for the first time and the accuracy boost can be up to 51.49\%. Without bells and whistles, \textsc{QDrop} establishes a new state of the art for PTQ.

****

ANCER: Anisotropic Certification via Sample-wise Volume Maximization

Francisco Eiras,Motasem Alfarra,Philip Torr,M. Pawan Kumar,Puneet K. Dokania,Bernard Ghanem,Adel Bibi

Randomized smoothing has recently emerged as an effective tool that enables certification of deep neural network classifiers at scale. All prior art on randomized smoothing has focused on isotropic ℓ_p certification, which has the advantage of yielding certificates that can be easily compared among isotropic methods via ℓ_p-norm radius. However, isotropic certification limits the region that can be certified around an input to worst-case adversaries, i.e., it cannot reason about other "close", potentially large, constant prediction safe regions. To alleviate this issue, (i) we theoretically extend the isotropic randomized smoothing ℓ_1 and ℓ_2 certificates to their generalized anisotropic counterparts following a simplified analysis. Moreover, (ii) we propose evaluation metrics allowing for the comparison of general certificates - a certificate is superior to another if it certifies a superset region - with the quantification of each certificate through the volume of the certified region. We introduce ANCER, a framework for obtaining anisotropic certificates for a given test set sample via volume maximization. We achieve it by generalizing memory-based certification of data-dependent classifiers. Our empirical results demonstrate that ANCER achieves state-of-the-art ℓ_1 and ℓ_2 certified accuracy on CIFAR-10 and ImageNet, while certifying larger regions in terms of volume, highlighting the benefits of moving away from isotropic analysis.

****

Data-Dependent Randomized Smoothing

Motasem Alfarra,Adel Bibi,Philip Torr,Bernard Ghanem

Randomized smoothing is a recent technique that achieves state-of-art performance in training certifiably robust deep neural networks. While the smoothing family of distributions is often connected to the choice of the norm used for certification, the parameters of these distributions are always set as global hyper parameters independent from the input data on which a network is certified. In this work, we revisit Gaussian randomized smoothing and show that the variance of the Gaussian distribution can be optimized at each input so as to maximize the certification radius for the construction of the smooth classifier. We also propose a simple memory-based approach to certifying the resultant smooth classifier. This new approach is generic, parameter-free, and easy to implement. In fact, we show that our data dependent framework can be seamlessly incorporated into 3 randomized smoothing approaches, leading to consistent improved certified accuracy. When this framework is used in the training routine of these approaches followed by a data dependent certification, we achieve 9% and 6% improvement over the certified accuracy of the strongest baseline for a radius of 0.5 on CIFAR10 and ImageNet.

****

You Mostly Walk Alone: Analyzing Feature Attribution in Trajectory Prediction

Osama Makansi,Julius Von Kügelgen,Francesco Locatello,Peter Vincent Gehler,Dominik Janzing,Thomas Brox,Bernhard Schölkopf

Predicting the future trajectory of a moving agent can be easy when the past trajectory continues smoothly but is challenging when complex interactions with other agents are involved. Recent deep learning approaches for trajectory prediction show promising performance and partially attribute this to successful reasoning about agent-agent interactions. However, it remains unclear which features such black-box models actually learn to use for making predictions. This paper proposes a procedure that quantifies the contributions of different cues to model performance based on a variant of Shapley values. Applying this procedure to state-of-the-art trajectory prediction methods on standard benchmark datasets shows that they are, in fact, unable to reason about interactions. Instead, the past trajectory of the target is the only feature used for predicting its future. For a task with richer social interaction patterns, on the other hand, the tested models do pick up such interactions to a certain extent, as quantified by our feature attribution method. We discuss the limits of the proposed method and its links to causality.

****

Rethinking Class-Prior Estimation for Positive-Unlabeled Learning

Yu Yao,Tongliang Liu,Bo Han,Mingming Gong,Gang Niu,Masashi Sugiyama,Dacheng Tao

Given only positive (P) and unlabeled (U) data, PU learning can train a binary classifier without any negative data. It has two building blocks: PU class-prior estimation (CPE) and PU classification; the latter has been well studied while the former has received less attention. Hitherto, the distributional-assumption-free CPE methods rely on a critical assumption that the support of the positive data distribution cannot be contained in the support of the negative data distribution. If this is violated, those CPE methods will systematically overestimate the class prior; it is even worse that we cannot verify the assumption based on the data. In this paper, we rethink CPE for PU learning—can we remove the assumption to make CPE always valid? We show an affirmative answer by proposing Regroup

ing CPE (ReCPE) that builds an auxiliary probability distribution such that the support of the positive data distribution is never contained in the support of the negative data distribution. ReCPE can work with any CPE method by treating it as the base method. Theoretically, ReCPE does not affect its base if the assumption already holds for the original probability distribution; otherwise, it reduces the positive bias of its base. Empirically, ReCPE improves all state-of-the-art CPE methods on various datasets, implying that the assumption has indeed been violated here.
****

JOINTLY LEARNING TOPIC SPECIFIC WORD AND DOCUMENT EMBEDDING
Farid Uddin,Zuping Zhang
Document embedding generally ignores underlying topics, which fails to capture polysemous terms that can mislead to improper thematic representation. Moreover, embedding a new document during the test process needs a complex and expensive inference method. Some models first learn word embeddings and later learn underlying topics using a clustering algorithm for document representation; those methods miss the mutual interaction between the two paradigms. To this point, we propose a novel document-embedding method by weighted averaging of jointly learning topic-specific word embeddings called TDE: Topical Document Embedding, which efficiently captures syntactic and semantic properties by utilizing three levels of knowledge -i.e., word, topic, and document. TDE obtains document vectors on the fly simultaneously during the jointly learning process of the topical word embeddings. Experiments demonstrate better topical word embeddings using document vector as a global context and better document classification results on the obtained document embeddings by the proposed method over the recent related models.
****

Thompson Sampling for (Combinatorial) Pure Exploration
Siwei Wang,Jun Zhu
Pure exploration plays an important role in online learning. Existing work mainly focuses on the UCB approach that uses confidence bounds of all the arms to decide which one is optimal. However, the UCB approach faces some challenges when looking for the best arm set under some specific combinatorial structures. It uses the sum of upper confidence bounds within arm set S to judge whether S is optimal. This sum can be much larger than the exact upper confidence bound of S, since the empirical means of different arms in S are independent. Because of this, the UCB approach requires much higher complexity than necessary. To deal with this challenge, we explore the idea of Thompson Sampling (TS) that uses independent random samples instead of the upper confidence bounds to make decisions, and design the first TS-based algorithm framework TS-Verify for (combinatorial) pure exploration. In TS-Verify, the sum of independent random samples within arm set S will not exceed the exact upper confidence bound of S with high probability. Hence it solves the above challange, and behaves better than existing UCB-based algorithms under the general combinatorial pure exploration setting. As for pure exploration of classic multi-armed bandit, we show that TS-Verify achieves an asymptotically optimal complexity upper bound.
****

SHINE: SHaring the INverse Estimate from the forward pass for bi-level optimization and implicit models
Zaccharie Ramzi,Florian Mannel,Shaojie Bai,Jean-Luc Starck,Philippe Ciuciu,Thomas Moreau
In recent years, implicit deep learning has emerged as a method to increase the depth of deep neural networks. While their training is memory-efficient, they are still significantly slower to train than their explicit counterparts. In Deep Equilibrium Models~(DEQs), the training is performed as a bi-level problem, and its computational complexity is partially driven by the iterative inversion of a huge Jacobian matrix. In this paper, we propose a novel strategy to tackle this computational bottleneck from which many bi-level problems suffer. The main idea is to use the quasi-Newton matrices from the forward pass to efficiently approximate the inverse Jacobian matrix in the direction needed for the gradient computation. We provide a theorem that motivates using our method with the original

forward algorithms. In addition, by modifying these forward algorithms, we further provide theoretical guarantees that our method asymptotically estimates the true implicit gradient. We empirically study this approach in many settings, ranging from hyperparameter optimization to large Multiscale DEQs applied to CIFAR and ImageNet. We show that it reduces the computational cost of the backward pass by up to two orders of magnitude. All this is achieved while retaining the excellent performance of the original models in hyperparameter optimization and on CIFAR, and giving encouraging and competitive results on ImageNet.

****

A stepped sampling method for video detection using LSTM

Dengshan Li

Artificial neural networks that simulate human achieves great successes. From the perspective of simulating human memory method, we propose a stepped sampler based on the “repeated input”. We repeatedly inputted data to the LSTM model stepwise in a batch. The stepped sampler is used to strengthen the ability of fusing the temporal information in LSTM. We tested the stepped sampler on the LSTM built-in in PyTorch. Compared with the traditional sampler of PyTorch, such as sequential sampler, batch sampler, the training loss of the proposed stepped sampler converges faster in the training of the model, and the training loss after convergence is more stable. Meanwhile, it can maintain a higher test accuracy. We quantified the algorithm of the stepped sampler. We assume that, the artificial neural networks have human-like characteristics, and human learning method could be used for machine learning. Our code will be available online soon.

****

Planckian jitter: enhancing the color quality of self-supervised visual representations

Simone Zini,Marco Buzzelli,Bart■omiej Twardowski,Joost van de weijer

Several recent works on self-supervised learning are trained by mapping different augmentations of the same image to the same feature representation. The set of used data augmentations is of crucial importance for the quality of the learned feature representation. We analyze how the traditionally used color jitter negatively impacts the quality of the color features in the learned feature representation. To address this problem, we replace this module with physics-based color augmentation, called Planckian jitter, which creates realistic variations in chromaticity, producing a model robust to llumination changes that can be commonly observed in real life, while maintaining the ability to discriminate the image content based on color information.

We further improve the performance by introducing a latent space combination of color-sensitive and non-color-sensitive features.

These are found to be complementary and the combination leads to large absolute performance gains over the default data augmentation on color classification tasks, including on Flowers-102 (+15%), Cub200 (+11%), VegFru (+15%), and T1K+ (+12%). Finally, we present a color sensitivity analysis to document the impact of different training methods on the model neurons and we show that the performance of the learned features is robust with respect to illuminant variations.

****

Stable cognitive maps for Path Integration emerge from fusing visual and proprioceptive sensors

Arnaud Fanthomme,Rémi Monasson

Spatial navigation in biological agents relies on the interplay between external (visual, olfactory, auditory, $\dots$) and proprioceptive (motor commands, linear and angular velocity, $\dots$) signals. How to combine and exploit these two streams of information, which vastly differ in terms of availability and reliability is a crucial issue. In the context of a new two--dimensional continuous environment we developed, we propose a direct-inverse model of environment dynamics to fuse image and action related signals, allowing reconstruction of the action relating the two successive images, as well as prediction of the new image from its current value and the action. The definition of those models naturally leads to the proposal of a minimalistic recurrent architecture, called Resetting Path Integrator (RPI), that can easily and reliably be trained to keep track of its

position relative to its starting point during a sequence of movements. RPI updates its internal state using the (possibly noisy) proprioceptive signal, and occasionally resets it when the image signal is present. Notably, the internal state of this minimal model exhibits strong correlation with position in the environment due to the direct-inverse models, is stable across long trajectories through resetting, and allows for disambiguation of visually confusing positions in the environment through integration of past movement, making it a prime candidate for a \textbf{cognitive map}. Our architecture is compared to state-of-the-art LSTM networks on identical tasks, and consistently shows better performance while also offering more interpretable internal dynamics and higher-quality representations.
****

Complex-valued deep learning with differential privacy
Alexander Ziller,Dmitrii Usynin,Moritz Knolle,Kerstin Hammernik,Daniel Rueckert,Georgios Kaissis

We present ζ-DP, an extension of differential privacy (DP) to complex-valued functions. After introducing the complex Gaussian mechanism, whose properties we characterise in terms of (ε, δ)-DP and Rényi-DP, we present ζ-DP stochastic gradient descent (ζ-DP-SGD), a variant of DP-SGD for training complex-valued neural networks. We experimentally evaluate ζ-DP-SGD on three complex-valued tasks, i.e. electrocardiogram classification, speech classification and magnetic resonance imaging (MRI) reconstruction. Moreover, we provide ζ-DP-SGD benchmarks for a large variety of complex-valued activation functions and on a complex-valued variant of the MNIST dataset. Our experiments demonstrate that DP training of complex-valued neural networks is possible with rigorous privacy guarantees and excellent utility.
****

Learning Efficient Online 3D Bin Packing on Packing Configuration Trees
Hang Zhao,Yang Yu,Kai Xu

Online 3D Bin Packing Problem (3D-BPP) has widespread applications in industrial automation and has aroused enthusiastic research interest recently. Existing methods usually solve the problem with limited resolution of spatial discretization, and/or cannot deal with complex practical constraints well. We propose to enhance the practical applicability of online 3D-BPP via learning on a novel hierarchical representation – packing configuration tree (PCT). PCT is a full-fledged description of the state and action space of bin packing which can support packing policy learning based on deep reinforcement learning (DRL). The size of the packing action space is proportional to the number of leaf nodes, making the DRL model easy to train and well-performing even with continuous solution space. During training, PCT expands based on heuristic rules, however, the DRL model learns a much more effective and robust packing policy than heuristic methods. Through extensive evaluation, we demonstrate that our method outperforms all existing online BPP methods and is versatile in terms of incorporating various practical constraints.
****

MT-GBM: A Multi-Task Gradient Boosting Machine with Shared Decision Trees
Zhenzhe Ying,Zhuoer Xu,LANQING XUE,Changhua Meng,Weiqiang Wang

Despite the success of deep learning in computer vision and natural language processing, Gradient Boosted Decision Tree (GBDT) is yet one of the most powerful tools for applications with tabular data such as e-commerce and FinTech. However, applying GBDT to multi-task learning is still a challenge. Unlike deep models that can jointly learn a shared latent representation across multiple tasks, GBDT can hardly learn a shared tree structure.

In this paper, we propose Multi-Task Gradient Boosting Machine (MT-GBM), a GBDT-based method for multi-task learning. The MT-GBM can find the shared tree structures and split branches according to multi-task losses. First, it assigns multiple outputs to each leaf node. Next, it computes the gradient corresponding to each output (task). Then, we also propose an algorithm to combine the gradients of all tasks and update the tree. Finally, we apply MT-GBM to LightGBM. Experiment s show that our MT-GBM improves the performance of the main task significantly, which means the proposed MT-GBM is efficient and effective.

****

There are free lunches

Zhuoran Xu,hao liu,bo dong

No-Free-Lunch Theorems state that the performance of all algorithms is the same when averaged over all possible tasks. It has been argued that the necessary conditions for NFL are too restrictive to be found in practice. There must be some information for a set of tasks that ensures some algorithms perform better than others. In this paper we propose a novel idea, "There are free lunches" (TAFL) Theorem, which states that some algorithms can achieve the best performance in all possible tasks, in the condition that tasks are given in a specific order. Furthermore, we point out that with the number of solved tasks increasing, the difficulty of solving a new task decreases. We also present an example to explain how to combine the proposed theorem and the existing supervised learning algorithms.

****

Towards Deployment-Efficient Reinforcement Learning: Lower Bound and Optimality

Jiawei Huang,Jinglin Chen,Li Zhao,Tao Qin,Nan Jiang,Tie-Yan Liu

Deployment efficiency is an important criterion for many real-world applications of reinforcement learning (RL). Despite the community's increasing interest, there lacks a formal theoretical formulation for the problem. In this paper, we propose such a formulation for deployment-efficient RL (DE-RL) from an ''optimization with constraints'' perspective: we are interested in exploring an MDP and obtaining a near-optimal policy within minimal \emph{deployment complexity}, whereas in each deployment the policy can sample a large batch of data. Using finite-horizon linear MDPs as a concrete structural model, we reveal the fundamental limit in achieving deployment efficiency by establishing information-theoretic lower bounds, and provide algorithms that achieve the optimal deployment efficiency. Moreover, our formulation for DE-RL is flexible and can serve as a building block for other practically relevant settings; we give ''Safe DE-RL'' and ''Sample-Efficient DE-RL'' as two examples, which may be worth future investigation.

****

Uncertainty Modeling for Out-of-Distribution Generalization

Xiaotong Li,Yongxing Dai,Yixiao Ge,Jun Liu,Ying Shan,LINGYU DUAN

Though remarkable progress has been achieved in various vision tasks, deep neural networks still suffer obvious performance degradation when tested in out-of-distribution scenarios. We argue that the feature statistics (mean and standard deviation), which carry the domain characteristics of the training data, can be properly manipulated to improve the generalization ability of deep learning models. Common methods often consider the feature statistics as deterministic values measured from the learned features and do not explicitly consider the uncertain statistics discrepancy caused by potential domain shifts during testing. In this paper, we improve the network generalization ability by modeling the uncertainty of domain shifts with synthesized feature statistics during training. Specifically, we hypothesize that the feature statistic, after considering the potential uncertainties, follows a multivariate Gaussian distribution. Hence, each feature statistic is no longer a deterministic value, but a probabilistic point with diverse distribution possibilities. With the uncertain feature statistics, the models can be trained to alleviate the domain perturbations and achieve better robustness against potential domain shifts. Our method can be readily integrated into networks without additional parameters. Extensive experiments demonstrate that our proposed method consistently improves the network generalization ability on multiple vision tasks, including image classification, semantic segmentation, and instance retrieval.

****

Online Adversarial Attacks

Andjela Mladenovic,Joey Bose,Hugo berard,William L. Hamilton,Simon Lacoste-Julien,Pascal Vincent,Gauthier Gidel

Adversarial attacks expose important vulnerabilities of deep learning models, ye

t little attention has been paid to settings where data arrives as a stream. In this paper, we formalize the online adversarial attack problem, emphasizing two key elements found in real-world use-cases: attackers must operate under partial knowledge of the target model, and the decisions made by the attacker are irrevocable since they operate on a transient data stream. We first rigorously analyze a deterministic variant of the online threat model by drawing parallels to the well-studied k-secretary problem in theoretical computer science and propose Virtual+, a simple yet practical online algorithm. Our main theoretical result shows Virtual+ yields provably the best competitive ratio over all single-threshold algorithms for $k<5$---extending the previous analysis of the k-secretary problem. We also introduce the \textit{stochastic k-secretary}---effectively reducing online blackbox transfer attacks to a k-secretary problem under noise---and prove theoretical bounds on the performance of Virtual+ adapted to this setting. Finally, we complement our theoretical results by conducting experiments on MNIST, CIFAR-10, and Imagenet classifiers, revealing the necessity of online algorithms in achieving near-optimal performance and also the rich interplay between attack strategies and online attack selection, enabling simple strategies like FGSM to outperform stronger adversaries.
****

Anytime Dense Prediction with Confidence Adaptivity

Zhuang Liu,Zhiqiu Xu,Hung-Ju Wang,Trevor Darrell,Evan Shelhamer

Anytime inference requires a model to make a progression of predictions which might be halted at any time. Prior research on anytime visual recognition has mostly focused on image classification.We propose the first unified and end-to-end approach for anytime dense prediction. A cascade of "exits" is attached to the model to make multiple predictions. We redesign the exits to account for the depth and spatial resolution of the features for each exit. To reduce total computation, and make full use of prior predictions, we develop a novel spatially adaptive approach to avoid further computation on regions where early predictions are already sufficiently confident. Our full method, named anytime dense prediction with confidence (ADP-C), achieves the same level of final accuracy, and meanwhile significantly reduces total computation. We evaluate our method on Cityscapes semantic segmentation and MPII human pose estimation: ADP-C enables anytime inference without sacrificing accuracy while also reducing the total FLOPs of its base models by 44.4% and 59.1%. We compare with anytime inference by deep equilibrium networks and feature-based stochastic sampling, showing that ADP-C dominates both across the accuracy-computation curve. Our code is available at https://github.com/liuzhuang13/anytime.
****

Declarative nets that are equilibrium models

Russell Tsuchida,Suk Yee Yong,Mohammad Ali Armin,Lars Petersson,Cheng Soon Ong

Implicit layers are computational modules that output the solution to some problem depending on the input and the layer parameters. Deep equilibrium models (DEQs) output a solution to a fixed point equation. Deep declarative networks (DDNs) solve an optimisation problem in their forward pass, an arguably more intuitive, interpretable problem than finding a fixed point. We show that solving a kernelised regularised maximum likelihood estimate as an inner problem in a DDN yields a large class of DEQ architectures. Our proof uses the exponential family in canonical form, and provides a closed-form expression for the DEQ parameters in terms of the kernel. The activation functions have interpretations in terms of the derivative of the log partition function. Building on existing literature, we interpret DEQs as fine-tuned, unrolled classical algorithms, giving an intuitive justification for why DEQ models are sensible. We use our theoretical result to devise an initialisation scheme for DEQs that allows them to solve kGLMs in their forward pass at initialisation. We empirically show that this initialisation scheme improves training stability and performance over random initialisation.
****

Adversarial Style Augmentation for Domain Generalized Urban-Scene Segmentation

Zhun Zhong,Yuyang Zhao,Gim Hee Lee,Nicu Sebe

In this paper, we consider the problem of domain generalization in semantic segm entation, 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.

****

EMFlow: Data Imputation in Latent Space via EM and Deep Flow Models

Qi Ma,Sujit K Ghosh

The presence of missing values within high-dimensional data is an ubiquitous problem for many applied sciences. A serious limitation of many available data mining and machine learning methods is their inability to handle partially missing values and so an integrated approach that combines imputation and model estimation is vital for down-stream analysis. A computationally fast algorithm, called EMFlow, is introduced that performs imputation in a latent space via an online version of Expectation-Maximization (EM) algorithm by using a normalizing flow (NF) model which maps the data space to a latent space. The proposed EMFlow algorithm is iterative, involving updating the parameters of online EM and NF alternatively. Extensive experimental results for high-dimensional multivariate and image datasets are presented to illustrate the superior performance of the EMFlow compared to a couple of recently available methods in terms of both predictive accuracy and speed of algorithmic convergence.

****

A Reduction-Based Framework for Conservative Bandits and Reinforcement Learning

Yunchang Yang,Tianhao Wu,Han Zhong,Evrard Garcelon,Matteo Pirotta,Alessandro Lazaric,Liwei Wang,Simon Shaolei Du

We study bandits and reinforcement learning (RL) subject to a conservative constraint where the agent is asked to perform at least as well as a given baseline policy. This setting is particular relevant in real-world domains including digital marketing, healthcare, production, finance, etc. In this paper, we present a reduction-based framework for conservative bandits and RL, in which our core technique is to calculate the necessary and sufficient budget obtained from running the baseline policy. For lower bounds, we improve the existing lower bound for conservative multi-armed bandits and obtain new lower bounds for conservative linear bandits, tabular RL and low-rank MDP, through a black-box reduction that turns a certain lower bound in the nonconservative setting into a new lower bound in the conservative setting. For upper bounds, in multi-armed bandits, linear bandits and tabular RL, our new upper bounds tighten or match existing ones with significantly simpler analyses. We also obtain a new upper bound for conservative low-rank MDP.

****

Beyond Target Networks: Improving Deep Q-learning with Functional Regularization

Alexandre Piché,Joseph Marino,Gian Maria Marconi,Christopher Pal,Mohammad Emtiyaz Khan

A majority of recent successes in deep Reinforcement Learning are based on minimization of square Bellman error. The training is often unstable due to a fast-changing target Q-values, and target networks are employed to stabilize by using an additional set of lagging parameters. Despite their advantages, target networks could inhibit the propagation of newly-encountered rewards which may ultimat ely slow down the training. In this work, we address this issue by augmenting the squared Bellman error with a functional regularizer. Unlike target networks', the regularization here is explicit which not only enables us to use up-to-date parameters but also control the regularization. This leads to a fast yet stable training method. Across a range of Atari environments, we demonstrate empirical improvements over target-network based methods in terms of both sample efficiency and performance. In summary, our approach provides a fast and stable alternative to replace the standard squared Bellman error.

****

Wisdom of Committees: An Overlooked Approach To Faster and More Accurate Models

Xiaofang Wang,Dan Kondratyuk,Eric Christiansen,Kris M. Kitani,Yair Movshovitz-Attias,Elad Eban

Committee-based models (ensembles or cascades) construct models by combining existing pre-trained ones. While ensembles and cascades are well-known techniques that were proposed before deep learning, they are not considered a core building block of deep model architectures and are rarely compared to in recent literature on developing efficient models. In this work, we go back to basics and conduct a comprehensive analysis of the efficiency of committee-based models. We find that even the most simplistic method for building committees from existing, independently pre-trained models can match or exceed the accuracy of state-of-the-art models while being drastically more efficient. These simple committee-based models also outperform sophisticated neural architecture search methods (e.g., BigNAS). These findings hold true for several tasks, including image classification, video classification, and semantic segmentation, and various architecture families, such as ViT, EfficientNet, ResNet, MobileNetV2, and X3D. Our results show that an EfficientNet cascade can achieve a 5.4x speedup over B7 and a ViT cascade can achieve a 2.3x speedup over ViT-L-384 while being equally accurate.

****

Unsupervised Discovery of Object Radiance Fields

Hong-Xing Yu,Leonidas Guibas,Jiajun Wu

We study the problem of inferring an object-centric scene representation from a single image, aiming to derive a representation that explains the image formation process, captures the scene's 3D nature, and is learned without supervision. Most existing methods on scene decomposition lack one or more of these characteristics, due to the fundamental challenge in integrating the complex 3D-to-2D image formation process into powerful inference schemes like deep networks. In this paper, we propose unsupervised discovery of Object Radiance Fields (uORF), integrating recent progresses in neural 3D scene representations and rendering with deep inference networks for unsupervised 3D scene decomposition. Trained on multi-view RGB images without annotations, uORF learns to decompose complex scenes with diverse, textured background from a single image. We show that uORF enables novel tasks, such as scene segmentation and editing in 3D, and it performs well on these tasks and on novel view synthesis on three datasets.

****

Learning Predictive, Online Approximations of Explanatory, Offline Algorithms

Mattson Thieme,Ammar Gilani,Han Liu

In this work, we introduce a general methodology for approximating offline algorithms in online settings. By encoding the behavior of offline algorithms in graphs, we train a multi-task learning model to simultaneously detect behavioral structures which have already occurred and predict those that may come next. We demonstrate the methodology on both synthetic data and historical stock market data, where the contrast between explanation and prediction is particularly stark. Taken together, our work represents the first general and end-to-end differentiable approach for generating online approximations of offline algorithms.

****

Gradient Step Denoiser for convergent Plug-and-Play

Samuel Hurault,Arthur Leclaire,Nicolas Papadakis

Plug-and-Play methods constitute a class of iterative algorithms for imaging problems where regularization is performed by an off-the-shelf denoiser. Although Plug-and-Play methods can lead to tremendous visual performance for various image problems, the few existing convergence guarantees are based on unrealistic (or suboptimal) hypotheses on the denoiser, or limited to strongly convex data terms. In this work, we propose a new type of Plug-and-Play methods, based on half-quadratic splitting, for which the denoiser is realized as a gradient descent step on a functional parameterized by a deep neural network. Exploiting convergence results for proximal gradient descent algorithms in the non-convex setting, we show that the proposed Plug-and-Play algorithm is a convergent iterative scheme that targets stationary points of an explicit global functional. Besides, experiments show that it is possible to learn such a deep denoiser while not compromising the performance in comparison to other state-of-the-art deep denoisers used in Plug-and-Play schemes. We apply our proximal gradient algorithm to various ill-posed inverse problems, e.g. deblurring, super-resolution and inpainting. For all these applications, numerical results empirically confirm the convergence results. Experiments also show that this new algorithm reaches state-of-the-art performance, both quantitatively and qualitatively.

****

SpecTRA: Spectral Transformer for Graph Representation Learning

Anson Bastos,Abhishek Nadgeri,Kuldeep Singh,Hiroki Kanezashi,Toyotaro Suzumura,Isaiah Onando Mulang'

Transformers have recently been applied in the more generic domain of graphs. For the same, researchers proposed various positional and structural encoding schemes to overcome the limitation of transformers in modeling the positional invariance in graphs and graph topology. Some of these encoding techniques use the spectrum of the graph. In addition to graph topology, graph signals could be multi-channeled and contain heterogeneous information. We argue that transformers cannot model multichannel signals inherently spread over the graph spectrum. To this end, we propose SpecTRA, a novel approach that induces a spectral module into the transformer architecture to enable decomposition of graph spectrum and selectively learn useful information akin to filtering in the frequency domain. Results on standard benchmark datasets show that the proposed method performs comparably or better than existing transformer and GNN based architectures.

****

Surrogate Gap Minimization Improves Sharpness-Aware Training

Juntang Zhuang,Boqing Gong,Liangzhe Yuan,Yin Cui,Hartwig Adam,Nicha C Dvornek,sekhar tatikonda,James s Duncan,Ting Liu

The recently proposed Sharpness-Aware Minimization (SAM) improves generalization by minimizing a perturbed loss defined as the maximum loss within a neighborhood in the parameter space. However, we show that both sharp and flat minima can have a low perturbed loss, implying that SAM does not always prefer flat minima. Instead, we define a surrogate gap, a measure equivalent to the dominant eigenvalue of Hessian at a local minimum when the radius of neighborhood (to derive the perturbed loss) is small. The surrogate gap is easy to compute and feasible for direct minimization during training. Based on the above observations, we propose Surrogate Gap Guided Sharpness-Aware Minimization (GSAM), a novel improvement over SAM with negligible computation overhead. Conceptually, GSAM consists of two steps: 1) a gradient descent like SAM to minimize the perturbed loss, and 2) an ascent step in the orthogonal direction (after gradient decomposition) to minimize the surrogate gap and yet not affect the perturbed loss. GSAM seeks a region with both small loss (by step 1) and low sharpness (by step 2), giving rise to a model with high generalization capabilities. Theoretically, we show the convergence of GSAM and provably better generalization than SAM.Empirically, GSAM consistently improves generalization (e.g., +3.2% over SAM and +5.4% over AdamW on ImageNet top-1 accuracy for ViT-B/32). Code is released at https://sites.google.com/view/gsam-iclr22/home

****

R4D: Utilizing Reference Objects for Long-Range Distance Estimation

Yingwei Li,Tiffany Chen,Maya Kabkab,Ruichi Yu,Longlong Jing,Yurong You,Hang Zhao

Estimating the distance of objects is a safety-critical task for autonomous driving. Focusing on short-range objects, existing methods and datasets neglect the equally important long-range objects. In this paper, we introduce a challenging and under-explored task, which we refer to as Long-Range Distance Estimation, as well as two datasets to validate new methods developed for this task. We then proposeR4D, the first framework to accurately estimate the distance of long-range objects by using references with known distances in the scene. Drawing inspiration from human perception, R4D builds a graph by connecting a target object to all references. An edge in the graph encodes the relative distance information between a pair of target and reference objects. An attention module is then used to weigh the importance of reference objects and combine them into one target object distance prediction. Experiments on the two proposed datasets demonstrate the effectiveness and robustness of R4D by showing significant improvements compared to existing baselines. We’re looking to make the proposed dataset, Waymo Open Dataset - Long-Range Labels, available publicly at waymo.com/open/download.

****

Understanding Dimensional Collapse in Contrastive Self-supervised Learning

Li Jing,Pascal Vincent,Yann LeCun,Yuandong Tian

Self-supervised visual representation learning aims to learn useful representations without relying on human annotations. Joint embedding approach bases on maximizing the agreement between embedding vectors from different views of the same image. Various methods have been proposed to solve the collapsing problem where all embedding vectors collapse to a trivial constant solution. Among these methods, contrastive learning prevents collapse via negative sample pairs. It has been shown that non-contrastive methods suffer from a lesser collapse problem of a different nature: dimensional collapse, whereby the embedding vectors end up spanning a lower-dimensional subspace instead of the entire available embedding space. Here, we show that dimensional collapse also happens in contrastive learning. In this paper, we shed light on the dynamics at play in contrastive learning that leads to dimensional collapse. Inspired by our theory, we propose a novel contrastive learning method, called DirectCLR, which directly optimizes the representation space without relying on a trainable projector. Experiments show that DirectCLR outperforms SimCLR with a trainable linear projector on ImageNet.

****

Contrastive Mutual Information Maximization for Binary Neural Networks

Yuzhang Shang,Dan Xu,Ziliang Zong,Liqiang Nie,Yan Yan

Neural network binarization accelerates deep models by quantizing their weights and activations into 1-bit. However, there is still a huge performance gap between Binary Neural Networks (BNNs) and their full-precision counterparts. As the quantization error caused by weights binarization has been reduced in earlier works, the activations binarization becomes the major obstacle for further improvement of the accuracy. In spite of studies about the full-precision networks highlighting the distributions of activations, few works study the distribution of the binary activations in BNNs. In this paper, we introduce mutual information as the metric to measure the information shared by the binary and the latent full-precision activations. Then we maximize the mutual information by establishing a contrastive learning framework while training BNNs. Specifically, the representation ability of the BNNs is greatly strengthened via pulling the positive pairs with binary and full-precision activations from the same input samples, as well as pushing negative pairs from different samples (the number of negative pairs can be exponentially large). This benefits the downstream tasks, not only classification but also segmentation and depth estimation, etc. The experimental results show that our method can be implemented as a pile-up module on existing state-of-the-art binarization methods and can remarkably improve the performance over them on CIFAR-10/100 and ImageNet, in addition to the good generalization ability on NYUD-v2.

****

Neural networks with trainable matrix activation functions

Yuwen Li,Zhengqi Liu,Ludmil Zikatanov

The training process of neural networks usually optimize weights and bias parameters of linear transformations, while nonlinear activation functions are pre-specified and fixed. This work develops a systematic approach to constructing matrix activation functions whose entries are generalized from ReLU. The activation i s based on matrix-vector multiplications using only scalar multiplications and comparisons. The proposed activation functions depend on parameters that are trained along with the weights and bias vectors. Neural networks based on this approach are simple and efficient and are shown to be robust in numerical experiments.

****

The hidden label-marginal biases of segmentation losses

Bingyuan Liu,Jose Dolz,Adrian Galdran,Riadh Kobbi,Ismail Ben Ayed

Most segmentation losses are arguably variants of the Cross-Entropy (CE) or Dice losses. In the abundant segmentation literature, there is no clear consensus as to which of these losses is a better choice, with varying performances for each across different benchmarks and applications. In this work, we develop a theoretical analysis that links these two types of losses, exposing their advantages and weaknesses. First, we provide a constrained-optimization perspective showing that CE and Dice share a much deeper connection than previously thought: They both decompose into label-marginal penalties and closely related ground-truth matching penalties. Then, we provide bound relationships and an information-theoretic analysis, which uncover hidden label-marginal biases: Dice has an intrinsic bias towards specific extremely imbalanced solutions, whereas CE implicitly encourages the ground-truth region proportions. Our theoretical results explain the wide experimental evidence in the medical-imaging literature, whereby Dice losses bring improvements for imbalanced segmentation. It also explains why CE dominates natural-image problems with diverse class proportions, in which case Dice might have difficulty adapting to different label-marginal distributions. Based on our theoretical analysis, we propose a principled and simple solution, which enables to control explicitly the label-marginal bias. Our loss integrates CE with explicit ${\cal L}_1$ regularization, which encourages label marginals to match target class proportions, thereby mitigating class imbalance but without losing generality. Comprehensive experiments and ablation studies over different losses and applications validate our theoretical analysis, as well as the effectiveness of our explicit label-marginal regularizers.

****

FedPara: Low-rank Hadamard Product for Communication-Efficient Federated Learning

Nam Hyeon-Woo,Moon Ye-Bin,Tae-Hyun Oh

In this work, we propose a communication-efficient parameterization, $\texttt{FedPara}$, for federated learning (FL) to overcome the burdens on frequent model uploads and downloads. Our method re-parameterizes weight parameters of layers using low-rank weights followed by the Hadamard product. Compared to the conventional low-rank parameterization, our $\texttt{FedPara}$ method is not restricted to low-rank constraints, and thereby it has a far larger capacity. This property enables to achieve comparable performance while requiring 3 to 10 times lower communication costs than the model with the original layers, which is not achievable by the traditional low-rank methods. The efficiency of our method can be further improved by combining with other efficient FL optimizers. In addition, we extend our method to a personalized FL application, $\texttt{pFedPara}$, which separates parameters into global and local ones. We show that $\texttt{pFedPara}$ outperforms competing personalized FL methods with more than three times fewer parameters.

****

Structural Causal Interpretation Theorem

Matej Zecevic,Devendra Singh Dhami,Constantin A. Rothkopf,Kristian Kersting

Human mental processes allow for qualitative reasoning about causality in terms of mechanistic relations of the variables of interest, which we argue are naturally described by structural causal model (SCM). Since interpretations are being derived from mental models, the same applies for SCM. By defining a metric space on SCM, we provide a theoretical perspective on the comparison of mental models and thereby conclude that interpretations can be used for guiding a learning system towards true causality. To this effect, we present a theoretical analysis from first principles that results in a human-readable interpretation scheme cons istent with the provided causality that we name structural causal interpretations (SCI). Going further, we prove that any existing neural induction method (NIM) is in fact interpretable. Our first experiment (E1) assesses the quality of such NIM-based SCI. In (E2) we observe evidence for our conjecture on improved sample-efficiency for SCI-based learning. After conducting a small user study, in (E3) we observe superiority in human-based over NIM-based SCI in support of our initial hypothesis.

****

RegionViT: Regional-to-Local Attention for Vision Transformers

Chun-Fu Chen,Rameswar Panda,Quanfu Fan

Vision transformer (ViT) has recently shown its strong capability in achieving comparable results to convolutional neural networks (CNNs) on image classification. However, vanilla ViT simply inherits the same architecture from the natural language processing directly, which is often not optimized for vision applications. Motivated by this, in this paper, we propose a new architecture that adopts the pyramid structure and employ novel regional-to-local attention rather than global self-attention in vision transformers. More specifically, our model first generates regional tokens and local tokens from an image with different patch sizes, where each regional token is associated with a set of local tokens based on the spatial location. The regional-to-local attention includes two steps: first, the regional self-attention extracts global information among all regional tokens and then the local self-attention exchanges the information among one regional token and the associated local tokens via self-attention. Therefore, even though local self-attention confines the scope in a local region but it can still receive global information.

Extensive experiments on four vision tasks, including image classification, object and keypoint detection, semantics segmentation and action recognition, show that our approach outperforms or is on par with state-of-the-art ViT variants including many concurrent works. Our source codes and models are available at \url{https://github.com/IBM/RegionViT}.

****

Quadtree Attention for Vision Transformers

Shitao Tang,Jiahui Zhang,Siyu Zhu,Ping Tan

Transformers have been successful in many vision tasks, thanks to their capability of capturing long-range dependency. However, their quadratic computational complexity poses a major obstacle for applying them to vision tasks requiring dense predictions, such as object detection, feature matching, stereo, etc. We introduce QuadTree Attention, which reduces the computational complexity from quadratic to linear. Our quadtree transformer builds token pyramids and computes attention in a coarse-to-fine manner. At each level, the top K patches with the highest attention scores are selected, such that at the next level, attention is only evaluated within the relevant regions corresponding to these top K patches. We demonstrate that quadtree attention achieves state-of-the-art performance in various vision tasks, e.g. with 4.0% improvement in feature matching on ScanNet, about 50% flops reduction in stereo matching, 0.4-1.5% improvement in top-1 accuracy on ImageNet classification, 1.2-1.8% improvement on COCO object detection, and 0.7-2.4% improvement on semantic segmentation over previous state-of-the-art transformers. The codes are available at https://github.com/Tangshitao/QuadtreeAttention.

****

Visual Correspondence Hallucination

Hugo Germain,Vincent Lepetit,Guillaume Bourmaud

Given a pair of partially overlapping source and target images and a keypoint in the source image, the keypoint's correspondent in the target image can be either visible, occluded or outside the field of view. Local feature matching methods are only able to identify the correspondent's location when it is visible, while humans can also hallucinate its location when it is occluded or outside the field of view through geometric reasoning. In this paper, we bridge this gap by training a network to output a peaked probability distribution over the correspondent's location, regardless of this correspondent being visible, occluded, or ou tside the field of view. We experimentally demonstrate that this network is indeed able to hallucinate correspondences on pairs of images captured in scenes that were not seen at training-time. We also apply this network to an absolute camera pose estimation problem and find it is significantly more robust than state-of-the-art local feature matching-based competitors.

****

Secure Distributed Training at Scale

Eduard Gorbunov,Alexander Borzunov,Michael Diskin,Max Ryabinin

Some of the hardest problems in deep learning can be solved via pooling together computational resources of many independent parties, as is the case for scientific collaborations and volunteer computing. Unfortunately, any single participant in such systems can jeopardize the entire training run by sending incorrect updates, whether deliberately or by mistake. Training in presence of such peers requires specialized distributed training algorithms with Byzantine tolerance. These algorithms often sacrifice efficiency by introducing redundant communication or passing all updates through a trusted server. As a result, it can be infeasible to apply such algorithms to large-scale distributed deep learning, where models can have billions of parameters. In this work, we propose a novel protocol for secure (Byzantine-tolerant) decentralized training that emphasizes communication efficiency. We rigorously analyze this protocol: in particular, we provide theoretical bounds for its resistance against Byzantine and Sybil attacks and show that it has a marginal communication overhead. To demonstrate its practical effectiveness, we conduct large-scale experiments on image classification and language modeling in presence of Byzantine attackers.

****

What’s Wrong with Deep Learning in Tree Search for Combinatorial Optimization

Maximilian Böther,Otto Kißig,Martin Taraz,Sarel Cohen,Karen Seidel,Tobias Friedrich

Combinatorial optimization lies at the core of many real-world problems. Especially since the rise of graph neural networks (GNNs), the deep learning community has been developing solvers that derive solutions to NP-hard problems by learning the problem-specific solution structure. However, reproducing the results of these publications proves to be difficult. We make three contributions. First, we present an open-source benchmark suite for the NP-hard Maximum Independent Set problem, in both its weighted and unweighted variants. The suite offers a unified interface to various state-of-the-art traditional and machine learning-based solvers. Second, using our benchmark suite, we conduct an in-depth analysis of the popular guided tree search algorithm by Li et al. [NeurIPS 2018], testing various configurations on small and large synthetic and real-world graphs. By re-implementing their algorithm with a focus on code quality and extensibility, we show that the graph convolution network used in the tree search does not learn a meaningful representation of the solution structure, and can in fact be replaced by random values. Instead, the tree search relies on algorithmic techniques like graph kernelization to find good solutions. Thus, the results from the original publication are not reproducible. Third, we extend the analysis to compare the tree search implementations to other solvers, showing that the classical algorithmic solvers often are faster, while providing solutions of similar quality. Additionally, we analyze a recent solver based on reinforcement learning and observe that for this solver, the GNN is responsible for the competitive solution quality.

****

Deep Attentive Variational Inference

Ifigeneia Apostolopoulou,Ian Char,Elan Rosenfeld,Artur Dubrawski

Stochastic Variational Inference is a powerful framework for learning large-scale probabilistic latent variable models. However, typical assumptions on the factorization or independence of the latent variables can substantially restrict its capacity for inference and generative modeling. A major line of active research aims at building more expressive variational models by designing deep hierarchies of interdependent latent variables. Although these models exhibit superior performance and enable richer latent representations, we show that they incur dim

inishing returns: adding more stochastic layers to an already very deep model yields small predictive improvement while substantially increasing the inference and training time. Moreover, the architecture for this class of models favors local interactions among the latent variables between neighboring layers when designing the conditioning factors of the involved distributions. This is the first work that proposes attention mechanisms to build more expressive variational distributions in deep probabilistic models by explicitly modeling both local and global interactions in the latent space. Specifically, we propose deep attentive variational autoencoder and test it on a variety of established datasets. We show it achieves state-of-the-art log-likelihoods while using fewer latent layers and requiring less training time than existing models. The proposed non-local inference reduces computational footprint by alleviating the need for deep hierarchies. Project code:
https://github.com/ifiaposto/Deep_Attentive_VI

****

ARTEMIS: Attention-based Retrieval with Text-Explicit Matching and Implicit Similarity

Ginger Delmas,Rafael S. Rezende,Gabriela Csurka,Diane Larlus

An intuitive way to search for images is to use queries composed of an example image and a complementary text. While the first provides rich and implicit context for the search, the latter explicitly calls for new traits, or specifies how some elements of the example image should be changed to retrieve the desired target image. Current approaches typically combine the features of each of the two elements of the query into a single representation, which can then be compared to the ones of the potential target images. Our work aims at shedding new light on the task by looking at it through the prism of two familiar and related frameworks: text-to-image and image-to-image retrieval. Taking inspiration from them, we exploit the specific relation of each query element with the targeted image and derive light-weight attention mechanisms which enable to mediate between the two complementary modalities. We validate our approach on several retrieval benchmarks, querying with images and their associated free-form text modifiers. Our method obtains state-of-the-art results without resorting to side information, multi-level features, heavy pre-training nor large architectures as in previous works.

****

Robust Imitation Learning from Corrupted Demonstrations

Liu Liu,Ziyang Tang,Lanqing Li,Dijun Luo

We consider offline Imitation Learning from corrupted demonstrations where a constant fraction of data can be noise or even arbitrary outliers. Classical approaches such as Behavior Cloning assumes that demonstrations are collected by an presumably optimal expert, hence
may fail drastically when learning from corrupted demonstrations. We propose a novel robust algorithm by minimizing a Median-of-Means (MOM) objective which guarantees the accurate estimation of policy, even in the presence of constant fraction of outliers.

Our theoretical analysis shows that our robust method in the corrupted setting enjoys nearly the same error scaling and sample complexity guarantees as the classical Behavior Cloning in the expert demonstration setting. Our experiments on continuous-control benchmarks validate that existing algorithms are fragile under corrupted demonstration while our method exhibits the predicted robustness and effectiveness.

****

Neural network architectures for disentangling the multimodal structure of data ensembles

M. Alex O. Vasilescu

We introduce neural network architectures that model the mechanism that generates data and address the difficult problem of disentangling the multimodal structure of data ensembles. We provide (i) an autoencoder-decoder architecture that implements the M-mode SVD and (ii) a generalized autoencoder that employs a kernel activation and implements the doubly nonlinear Kernel-MPCA. The neural netwo rk projection architecture decomposes an unlabeled data given an estimated forward model and a set of observations that constrain the solution set.

****

Mistill: Distilling Distributed Network Protocols from Examples

Patrick Krämer,Johannes Zerwas,Andreas Blenk

New applications and use-cases in data center networks require the design of Traffic Engineering (TE) algorithms that account for application-specific traffic patterns. TE makes forwarding decisions from the global state of the network. Thus, new TE algorithms require the design and implementation of effective information exchange and efficient algorithms to compute forwarding decisions. This is a challenging and labor and time-intensive process. To automate and simplify this process, we propose MISTILL. MISTILL distills the forwarding behavior of TE policies from exemplary forwarding decisions into a Neural Network. MISTILL learns which network devices must exchange state with each other, how to process local state to send it over the network, and how to map the exchanged state into forwarding decisions. We show the ability of MISTILL to learn distributed protocols with three examples and verify their performance in simulations. We show that the learned protocols closely implement the desired policies.

****

A General Theory of Relativity in Reinforcement Learning

Lei Han,Cheng Zhou,Yizheng Zhang

We propose a new general theory measuring the relativity between two arbitrary Markov Decision Processes (MDPs) from the perspective of reinforcement learning (RL). Considering two MDPs, tasks such as policy transfer, dynamics modeling, environment design, and simulation to reality (sim2real), etc., are all closely related. The proposed theory deeply investigates the connection between any two cumulative expected returns defined on different policies and environment dynamics, and the theoretical results suggest two new general algorithms referred to as Relative Policy Optimization (RPO) and Relative Transition Optimization (RTO), which can offer fast policy transfer and dynamics modeling. RPO updates the policy using the \emph{relative policy gradient} to transfer the policy evaluated in one environment to maximize the return in another, while RTO updates the parameterized dynamics model (if there exists) using the \emph{relative transition gradient} to reduce the gap between the dynamics of the two environments. Then, integrating the two algorithms offers the complete algorithm Relative Policy-Transition Optimization (RPTO), in which the policy interacts with the two environments simultaneously, such that data collections from the two environments, policy and transition updates are all completed in a closed loop to form a principled learning framework for policy transfer. We demonstrate the effectiveness of RPO, RTO and RPTO in the OpenAI gym's classic control tasks by creating policy transfer problems.

****

Trivial or Impossible --- dichotomous data difficulty masks model differences (on ImageNet and beyond)

Kristof Meding,Luca M. Schulze Buschoff,Robert Geirhos,Felix A. Wichmann

"The power of a generalization system follows directly from its biases" (Mitchell 1980). Today, CNNs are incredibly powerful generalisation systems---but to what degree have we understood how their inductive bias influences model decisions? We here attempt to disentangle the various aspects that determine how a model decides. In particular, we ask: what makes one model decide differently from another? In a meticulously controlled setting, we find that (1.) irrespective of the network architecture or objective (e.g. self-supervised, semi-supervised, vision transformers, recurrent models) all models end up with a similar decision boundary. (2.) To understand these findings, we analysed model decisions on the ImageNet validation set from epoch to epoch and image by image. We find that the ImageNet validation set, among others, suffers from dichotomous data difficulty (DDD): For the range of investigated models and their accuracies, it is dominated by 46.0% "trivial" and 11.5% "impossible" images (beyond label errors). Only 42.5% of the images could possibly be responsible for the differences between two models' decision boundaries. (3.) Only removing the "impossible" and "trivial" im ages allows us to see pronounced differences between models. (4.) Humans are highly accurate at predicting which images are "trivial" and "impossible" for CNNs (81.4%). This implies that in future comparisons of brains, machines and behaviour, much may be gained from investigating the decisive role of images and the distribution of their difficulties.

****

Beyond Object Recognition: A New Benchmark towards Object Concept Learning
Yong-Lu Li,Yue Xu,Xinyu Xu,Xiaohan Mao,Yuan Yao,Siqi Liu,Cewu Lu
Understanding objects is a central building block of artificial intelligence, especially for embodied AI. Even though object recognition excels with deep learning, current machines still struggle to learn higher-level knowledge, e.g., what attributes does an object have, what can we do with an object. In this work, we propose a challenging Object Concept Learning (OCL) task to push the envelope of object understanding. It requires machines to reason out object affordances and simultaneously give the reason: what attributes make an object possesses these affordances. To support OCL, we build a densely annotated knowledge base including extensive labels for three levels of object concept: categories, attributes, and affordances, together with their causal relations. By analyzing the causal structure of OCL, we present a strong baseline, Object Concept Reasoning Network (OCRN). It leverages causal intervention and concept instantiation to infer the three levels following their causal relations. In extensive experiments, OCRN effectively infers the object knowledge while follows the causalities well. Our data and code will be publicly available.

****

A Scaling Law for Syn-to-Real Transfer: How Much Is Your Pre-training Effective?
Hiroaki Mikami,Kenji Fukumizu,Shogo Murai,Shuji Suzuki,Yuta Kikuchi,Taiji Suzuki,Shin-ichi Maeda,Kohei Hayashi
Synthetic-to-real transfer learning is a framework in which a synthetically generated dataset is used to pre-train a model to improve its performance on real vision tasks. The most significant advantage of using synthetic images is that the ground-truth labels are automatically available, enabling unlimited data size expansion without human cost. However, synthetic data may have a huge domain gap, in which case increasing the data size does not improve the performance. How can we know that? In this study, we derive a simple scaling law that predicts the performance from the amount of pre-training data. By estimating the parameters of the law, we can judge whether we should increase the data or change the setting of image synthesis. Further, we analyze the theory of transfer learning by considering learning dynamics and confirm that the derived generalization bound is compatible with our empirical findings. We empirically validated our scaling law on various experimental settings of benchmark tasks, model sizes, and complexities of synthetic images.

****

Group equivariant neural posterior estimation
Maximilian Dax,Stephen R Green,Jonathan Gair,Michael Deistler,Bernhard Schölkopf,Jakob H. Macke
Simulation-based inference with conditional neural density estimators is a powerful approach to solving inverse problems in science. However, these methods typically treat the underlying forward model as a black box, with no way to exploit geometric properties such as equivariances. Equivariances are common in scientific models, however integrating them directly into expressive inference networks (such as normalizing flows) is not straightforward. We here describe an alternative method to incorporate equivariances under joint transformations of parameters and data. Our method---called group equivariant neural posterior estimation (GNPE)---is based on self-consistently standardizing the "pose" of the data while estimating the posterior over parameters. It is architecture-independent, and applies both to exact and approximate equivariances. As a real-world application, we use GNPE for amortized inference of astrophysical binary black hole systems from gravitational-wave observations. We show that GNPE achieves state-of-the-art accuracy while reducing inference times by three orders of magnitude.

****

Fast Differentiable Matrix Square Root
Yue Song,Nicu Sebe,Wei Wang
Computing the matrix square root or its inverse in a differentiable manner is important in a variety of computer vision tasks. Previous methods either adopt the Singular Value Decomposition (SVD) to explicitly factorize the matrix or use the Newton-Schulz iteration (NS iteration) to derive the approximate solution. However, both methods are not computationally efficient enough in either the forward pass or in the backward pass. In this paper, we propose two more efficient variants to compute the differentiable matrix square root. For the forward propagation, one method is to use Matrix Taylor Polynomial (MTP), and the other method is to use Matrix Pad\'e Approximants (MPA). The backward gradient is computed by iteratively solving the continuous-time Lyapunov equation using the matrix sign function. Both methods yield considerable speed-up compared with the SVD or the Newton-Schulz iteration. Experimental results on the de-correlated batch normalization and second-order vision transformer demonstrate that our methods can also achieve competitive and even slightly better performances. The code is available at \href{https://github.com/KingJamesSong/FastDifferentiableMatSqrt}{https://github.com/KingJamesSong/FastDifferentiableMatSqrt}.
****
SQuant: On-the-Fly Data-Free Quantization via Diagonal Hessian Approximation
Cong Guo,Yuxian Qiu,Jingwen Leng,Xiaotian Gao,Chen Zhang,Yunxin Liu,Fan Yang,Yuhao Zhu,Minyi Guo
Quantization of deep neural networks (DNN) has been proven effective for compressing and accelerating DNN models. Data-free quantization (DFQ) is a promising approach without the original datasets under privacy-sensitive and confidential scenarios. However, current DFQ solutions degrade accuracy, need synthetic data to calibrate networks, and are time-consuming and costly. This paper proposes an on-the-fly DFQ framework with sub-second quantization time, called SQuant, which can quantize networks on inference-only devices with low computation and memory requirements. With the theoretical analysis of the second-order information of DNN task loss, we decompose and approximate the Hessian-based optimization objective into three diagonal sub-items, which have different areas corresponding to three dimensions of weight tensor: element-wise, kernel-wise, and output channel-wise. Then, we progressively compose sub-items and propose a novel data-free optimization objective in the discrete domain, minimizing Constrained Absolute Sum of Error (or CASE in short), which surprisingly does not need any dataset and is even not aware of network architecture. We also design an efficient algorithm without back-propagation to further reduce the computation complexity of the objective solver. Finally, without fine-tuning and synthetic datasets, SQuant accelerates the data-free quantization process to a sub-second level with >30% accuracy improvement over the existing data-free post-training quantization works, with the evaluated models under 4-bit quantization. We have open-sourced the SQuant framework at https://github.com/clevercool/SQuant.
****
Neural Variational Dropout Processes
Insu Jeon,Youngjin Park,Gunhee Kim
Learning to infer the conditional posterior model is a key step for robust meta-learning. This paper presents a new Bayesian meta-learning approach called Neural Variational Dropout Processes (NVDPs). NVDPs model the conditional posterior distribution based on a task-specific dropout; a low-rank product of Bernoulli experts meta-model is utilized for a memory-efficient mapping of dropout rates from a few observed contexts. It allows for a quick reconfiguration of a globally learned and shared neural network for new tasks in multi-task few-shot learning. In addition, NVDPs utilize a novel prior conditioned on the whole task data to optimize the conditional dropout posterior in the amortized variational inference. Surprisingly, this enables the robust approximation of task-specific dropout rates that can deal with a wide range of functional ambiguities and uncertainties. We compared the proposed method with other meta-learning approaches in the few-shot learning tasks such as 1D stochastic regression, image inpainting, and classification. The results show the excellent performance of NVDPs.

****

Towards Better Understanding and Better Generalization of Low-shot Classification in Histology Images with Contrastive Learning

Jiawei Yang,Hanbo Chen,Jiangpeng Yan,Xiaoyu Chen,Jianhua Yao

Few-shot learning is an established topic in natural images for years, but few work is attended to histology images, which is of high clinical value since well-labeled datasets and rare abnormal samples are expensive to collect. Here, we facilitate the study of few-shot learning in histology images by setting up three cross-domain tasks that simulate real clinics problems. To enable label-efficient learning and better generalizability, we propose to incorporate contrastive learning (CL) with latent augmentation (LA) to build a few-shot system. CL learns useful representations without manual labels, while LA transfers semantic variations of the base dataset in an unsupervised way. These two components fully exploit unlabeled training data and can scale gracefully to other label-hungry problems. In experiments, we find i) models learned by CL generalize better than supervised learning for histology images in unseen classes, and ii) LA brings consistent gains over baselines. Prior studies of self-supervised learning mainly focus on ImageNet-like images, which only present a dominant object in their centers. Recent attention has been paid to images with multi-objects and multi-textures. Histology images are a natural choice for such a study. We show the superiority of CL over supervised learning in terms of generalization for such data and provide our empirical understanding for this observation. The findings in this work could contribute to understanding how the model generalizes in the context of both representation learning and histological image analysis. Code is available.

****

Distilling GANs with Style-Mixed Triplets for X2I Translation with Limited Data

Yaxing Wang,Joost van de weijer,Lu Yu,SHANGLING JUI

Conditional image synthesis is an integral part of many X2I translation systems, including image-to-image, text-to-image and audio-to-image translation systems. Training these large systems generally requires huge amounts of training data. Therefore, we investigate knowledge distillation to transfer knowledge from a high-quality unconditioned generative model (e.g., StyleGAN) to a conditioned synthetic image generation modules in a variety of systems. To initialize the conditional and reference branch (from a unconditional GAN) we exploit the style mixing characteristics of high-quality GANs to generate an infinite supply of style-mixed triplets to perform the knowledge distillation. Extensive experimental results in a number of image generation tasks (i.e., image-to-image, semantic segmentation-to-image, text-to-image and audio-to-image) demonstrate qualitatively and quantitatively that our method successfully transfers knowledge to the synthetic image generation modules, resulting in more realistic images than previous methods as confirmed by a significant drop in the FID.

****

Tabular Data Imputation: Choose KNN over Deep Learning

Florian Lalande,Kenji Doya

As databases are ubiquitous nowadays, missing values constitute a pervasive problem for data analysis. Over the last 70 years, various imputation algorithms for tabular data have been developed and shown useful at estimating missing values. Besides, recent infatuations for Artificial Neural Networks have led to the development of complex and powerful algorithms for data imputation.

This study is the first to compare state-of-the-art deep-learning models with the well-established KNN algorithm (1951). By using real-world and generated datasets in various missing data scenarios, we claim that the good old KNN algorithm is still competitive (nay better) than powerful deep-learning algorithms for tabular data imputation.

This work advocates for an appropriate and reasonable use of machine learning, in a world where overconsumption, performances and rapidity unfortunately often prevails over sustainability and common sense.

****

Handling Distribution Shifts on Graphs: An Invariance Perspective

Qitian Wu,Hengrui Zhang,Junchi Yan,David Wipf

There is increasing evidence suggesting neural networks' sensitivity to distribution shifts, so that research on out-of-distribution (OOD) generalization comes into the spotlight. Nonetheless, current endeavors mostly focus on Euclidean data, and its formulation for graph-structured data is not clear and remains under-explored, given two-fold fundamental challenges: 1) the inter-connection among nodes in one graph, which induces non-IID generation of data points even under the same environment, and 2) the structural information in the input graph, which is also informative for prediction. In this paper, we formulate the OOD problem on graphs and develop a new invariant learning approach, Explore-to-Extrapolate Risk Minimization (EERM), that facilitates graph neural networks to leverage invariance principles for prediction. EERM resorts to multiple context explorers (specified as graph structure editers in our case) that are adversarially trained to maximize the variance of risks from multiple virtual environments. Such a design enables the model to extrapolate from a single observed environment which is the common case for node-level prediction. We prove the validity of our method by theoretically showing its guarantee of a valid OOD solution and further demonstrate its power on various real-world datasets for handling distribution shifts from artificial spurious features, cross-domain transfers and dynamic graph evolution.
****

Automatic Loss Function Search for Predict-Then-Optimize Problems with Strong Ranking Property

Boshi Wang,Jialin Yi,Hang Dong,Bo Qiao,Chuan Luo,Qingwei Lin

Combinatorial optimization problems with parameters to be predicted from side information are commonly seen in a variety of problems during the paradigm shift from reactive decision making to proactive decision making. Due to the misalignment between the continuous prediction results and the discrete decisions in optimization problems, it is hard to achieve a satisfactory prediction result with the ordinary l_2 loss in the prediction phase. To properly connect the prediction loss with the optimization goal, in this paper we propose a total group preorder (TGP) loss and its differential version called approximated total group preorder (ATGP) loss for predict-then-optimize (PTO) problems with strong ranking property. These new losses are provably more robust than the usual l_2 loss in a linear regression setting and have great potential to extend to other settings. We also propose an automatic searching algorithm that adapts the ATGP loss to PTO problems with different combinatorial structures. Extensive experiments on the ranking problem, the knapsack problem, and the shortest path problem have demonstrated that our proposed method can achieve a significant performance compared to the other methods designed for PTO problems.
****

Task Relatedness-Based Generalization Bounds for Meta Learning

Jiechao Guan,Zhiwu Lu

Supposing the n training tasks and the new task are sampled from the same environment, traditional meta learning theory derives an error bound on the expected loss over the new task in terms of the empirical training loss, uniformly over the set of all hypothesis spaces. However, there is still little research on how the relatedness of these tasks can affect the full utilization of all mn training data (with m examples per task). In this paper, we propose to address this problem by defining a new notion of task relatedness according to the existence of the bijective transformation between two tasks. A novel generalization bound of $\mathcal{O}(\frac{1}{\sqrt{mn}})$ for meta learning is thus derived by exploiting the proposed task relatedness. Moreover, when investigating a special branch of meta learning that involves representation learning with deep neural networks, we establish spectrally-normalized bounds for both classification and regression problems. Finally, we demonstrate that the relatedness requirement between two tasks is satisfied when the sample space possesses the completeness and separability properties, validating the rationality and applicability of our proposed task-relatedness measure.
****

Let Your Heart Speak in its Mother Tongue: Multilingual Captioning of Cardiac Si gnals

Dani Kiyasseh,Tingting Zhu,David A. Clifton

Cardiac signals convey a significant amount of information about the health status of a patient. Upon recording these signals, cardiologists are expected to manually generate an accompanying report to share with physicians and patients. Generating these reports, however, can be time-consuming and error-prone, while also exhibiting a high degree of intra- and inter-physician variability. To address this, we design a neural, multilingual, cardiac signal captioning framework. In the process, we propose a discriminative multilingual representation learning method, RTLP, which randomly replaces tokens with those from a different language and tasks a network with identifying the language of all tokens. We show that RTLP performs on par with state-of-the-art pre-training methods such as MLM and MARGE, while generating more clinically accurate reports than MLM. We also show that, with RTLP, multilingual fine-tuning can be preferable to its monolingual counterpart, a phenomenon we refer to as the \textit{blessing of multilinguality}.

****

Generalized Demographic Parity for Group Fairness

Zhimeng Jiang,Xiaotian Han,Chao Fan,Fan Yang,Ali Mostafavi,Xia Hu

This work aims to generalize demographic parity to continuous sensitive attributes while preserving tractable computation. Current fairness metrics for continuous sensitive attributes largely rely on intractable statistical independence between variables, such as Hirschfeld-Gebelein-Renyi (HGR) and mutual information. Statistical fairness metrics estimation relying on either tractable bounds or neural network approximation, however, are not sufficiently trustful to rank algorithms prediction bias due to lack of estimation accuracy guarantee.

To make fairness metrics trustable, we propose \textit{\underline{G}eneralized \underline{D}emographic \underline{P}arity} (GDP), a group fairness metric for continuous and discrete attributes. We show the understanding of GDP from the probability perspective and theoretically reveal the connection between GDP regularizer and adversarial debiasing. To estimate GDP, we adopt hard and soft group strategies via the one-hot or the soft group indicator, representing the membership of each sample in different groups of the sensitive attribute. We provably and numerically show that the soft group strategy achieves a faster estimation error convergence rate. Experiments show the better bias mitigation performance of GDP regularizer, compared with adversarial debiasing, for regression and classification tasks in tabular and graph benchmarks.

****

Closed-form Sample Probing for Learning Generative Models in Zero-shot Learning

Samet Cetin,Orhun Bu■ra Baran,Ramazan Gokberk Cinbis

Generative model based approaches have led to significant advances in zero-shot learning (ZSL) over the past few years. These approaches typically aim to learn a conditional generator that synthesizes training samples of classes conditioned on class definitions. The final zero-shot learning model is then obtained by training a supervised classification model over the real and/or synthesized training samples of seen and unseen classes, combined. Therefore, naturally, the generative model needs to produce not only relevant samples, but also those that are sufficiently rich for classifier training purposes, which is handled by various heuristics in existing works. In this paper, we introduce a principled approach for training generative models {\em directly} for training data generation purposes. Our main observation is that the use of closed-form models opens doors to end-to-end training thanks to the differentiability of the solvers. In our approach, at each generative model update step, we fit a task-specific closed-form ZSL model from generated samples, and measure its loss on novel samples all within the compute graph, a procedure that we refer to as {\em sample probing}. In this manner, the generator receives feedback directly based on the value of its samples for model training purposes. Our experimental results show that the proposed sample probing approach improves the ZSL results even when integrated into state-of-the-art generative models.

****

DKM: Differentiable k-Means Clustering Layer for Neural Network Compression
Minsik Cho,Keivan Alizadeh-Vahid,Saurabh Adya,Mohammad Rastegari
Deep neural network (DNN) model compression for efficient on-device inference is becoming increasingly important to reduce memory requirements and keep user data on-device. To this end, we propose a novel differentiable k-means clustering layer (DKM) and its application to train-time weight clustering-based DNN model compression. DKM casts k-means clustering as an attention problem and enables joint optimization of the DNN parameters and clustering centroids. Unlike prior works that rely on additional regularizers and parameters, DKM-based compression keeps the original loss function and model architecture fixed. We evaluated DKM-based compression on various DNN models for computer vision and natural language processing (NLP) tasks. Our results demonstrate that DKM delivers superior compression and accuracy trade-off on ImageNet1k and GLUE benchmarks. For example, DKM-based compression can offer 74.5% top-1 ImageNet1k accuracy on ResNet50 DNN model with 3.3MB model size (29.4x model compression factor). For MobileNet-v1, which is a challenging DNN to compress, DKM delivers 63.9% top-1 ImageNet1k accuracy with 0.72 MB model size (22.4x model compression factor). This result is 6.8% higher top-1accuracy and 33% relatively smaller model size than the current state-of-the-art DNN compression algorithms. Additionally, DKM enables compression of DistilBERT model by 11.8x with minimal (1.1%) accuracy loss on GLUE NLP benchmarks.
****

Modeling Bounded Rationality in Multi-Agent Simulations Using Rationally Inattentive Reinforcement Learning
Tong Mu,Stephan Zheng,Alexander R Trott
Multi-agent reinforcement learning (MARL) is a powerful framework for studying emergent behavior in complex agent-based simulations. However, RL agents are often assumed to be rational and behave optimally, which does not fully reflect human behavior. Here, we study more human-like RL agents which incorporate an established model of human-irrationality, the Rational Inattention (RI) model. RI models the cost of cognitive information processing using mutual information. Our RIRL framework generalizes and is more flexible than prior work by allowing for multi-timestep dynamics and information channels with heterogeneous processing costs. We evaluate RIRL in Principal-Agent (specifically manager-employee relations) problem settings of varying complexity where RI models information asymmetry (e.g. it may be costly for the manager to observe certain information about the employees). We show that using RIRL yields a rich spectrum of new equilibrium behaviors that differ from those found under rational assumptions. For instance, some forms of a Principal's inattention can increase Agent welfare due to increased compensation, while other forms of inattention can decrease Agent welfare by encouraging extra work effort. Additionally, new strategies emerge compared to those under rationality assumptions, e.g., Agents are incentivized to misrepresent their ability. These results suggest RIRL is a powerful tool towards building AI agents that can mimic real human behavior.
****

Task-aware Privacy Preservation for Multi-dimensional Data
Jiangnan Cheng,Ao Tang,Sandeep P. Chinchali
Local differential privacy (LDP), a state-of-the-art technique for privacy preservation, has been successfully deployed in a few real-world applications. In the future, LDP can be adopted to anonymize richer user data attributes that will be input to more sophisticated machine learning (ML) tasks. However, today's LDP approaches are largely task-agnostic and often lead to sub-optimal performance - they will simply inject noise to all data attributes according to a given privacy budget, regardless of what features are most relevant for an ultimate task. In this paper, we address how to significantly improve the ultimate task performance for multi-dimensional user data by considering a task-aware privacy preservation problem. The key idea is to use an encoder-decoder framework to learn (and anonymize) a task-relevant latent representation of user data, which gives an analytical near-optimal solution for a linear setting with mean-squared error (MSE) task loss. We also provide an approximate solution through a learning algorith

m for general nonlinear cases. Extensive experiments demonstrate that our task-aware approach significantly improves ultimate task accuracy compared to a standard benchmark LDP approach while guaranteeing the same level of privacy.

****

Adapting Stepsizes by Momentumized Gradients Improves Optimization and Generalization

Yizhou Wang,Yue Kang,Can Qin,Huan Wang,Yi Xu,Yulun Zhang,Yun Fu

Adaptive gradient methods, such as Adam, have achieved tremendous success in machine learning. Scaling gradients by square roots of the running averages of squared past gradients, such methods are able to attain rapid training of modern deep neural networks. Nevertheless, they are observed to generalize worse than stochastic gradient descent (SGD) and tend to be trapped in local minima at an early stage during training. Intriguingly, we discover that substituting the gradient in the second moment estimation term with the momentumized version in Adam can well solve the issues. The intuition is that gradient with momentum contains more accurate directional information and therefore its second moment estimation is a better choice for scaling than that of the raw gradient. Thereby we propose AdaMomentum as a new optimizer reaching the goal of training fast while generalizing better. We further develop a theory to back up the improvement in optimization and generalization and provide convergence guarantees under both convex and nonconvex settings. Extensive experiments on a wide range of tasks and models demonstrate that AdaMomentum exhibits state-of-the-art performance consistently. The source code is available at https://anonymous.4open.science/r/AdaMomentum_experiments-6D9B.

****

Towards Generic Interface for Human-Neural Network Knowledge Exchange

Yunhao Ge,Yao Xiao,Zhi Xu,Linwei Li,Ziyan Wu,Laurent Itti

Neural Networks (NN) outperform humans in multiple domains. Yet they suffer from a lack of transparency and interpretability, which hinders intuitive and effective human interactions with them. Especially when NN makes mistakes, humans can hardly locate the reason for the error, and correcting it is even harder. While recent advances in explainable AI have substantially improved the explainability of NNs, effective knowledge exchange between humans and NNs is still under-explored. To fill this gap, we propose Human-NN-Interface (HNI), a framework using a structural representation of visual concepts as a ”language” for humans and NN to communicate, interact, and exchange knowledge. Take image classification as an example, HNI visualizes the reasoning logic of a NN with class-specific Structural Concept Graphs (c-SCG), which are human-interpretable. On the other hand, humans can effectively provide feedback and guidance to the NN by modifying the c-SCG, and transferring the knowledge back to NN through HNI. We demonstrate the efficacy of HNI with image classification tasks and 3 different types of interactions: (1) Explaining the reasoning logic of NNs so humans can intuitively identify and locate errors of NN; (2) human users can correct the errors and improve NN’s performance by modifying the c-SCG and distilling the knowledge back to the original NN; (3) human users can intuitively guide NN and provide a new solution for zero-shot learning.

****

Fixed Neural Network Steganography: Train the images, not the network

Varsha Kishore,Xiangyu Chen,Yan Wang,Boyi Li,Kilian Q Weinberger

Recent attempts at image steganography make use of advances in deep learning to train an encoder-decoder network pair to hide and retrieve secret messages in images. These methods are able to hide large amounts of data, but they also incur high decoding error rates (around 20%). In this paper, we propose a novel algorithm for steganography that takes advantage of the fact that neural networks are sensitive to tiny perturbations. Our method, Fixed Neural Network Steganography (FNNS), yields significantly lower error rates when compared to prior state-of-the-art methods and achieves 0% error reliably for hiding up to 3 bits per pixel (bpp) of secret information in images. FNNS also successfully evades existing statistical steganalysis systems and can be modified to evade neural steganalysis systems as well. Recovering every bit correctly, up to 3 bpp, enables novel appl ications that requires encryption. We introduce one specific use case for facilitating anonymized and safe image sharing. Our code is available at https://github.com/varshakishore/FNNS.

****

Finding General Equilibria in Many-Agent Economic Simulations using Deep Reinforcement Learning

Michael Curry,Alexander R Trott,Soham Phade,Yu Bai,Stephan Zheng

Real economies can be seen as a sequential imperfect-information game with many heterogeneous, interacting strategic agents of various agent types, such as consumers, firms, and governments. Dynamic general equilibrium models are common economic tools to model the economic activity, interactions, and outcomes in such systems. However, existing analytical and computational methods struggle to find explicit equilibria when all agents are strategic and interact, while joint learning is unstable and challenging. Amongst others, a key reason is that the actions of one economic agent may change the reward function of another agent, e.g., a consumer's expendable income changes when firms change prices or governments change taxes. We show that multi-agent deep reinforcement learning (RL) can discover stable solutions that are ϵ-Nash equilibria for a meta-game over agent types, in economic simulations with many agents, through the use of structured learning curricula and efficient GPU-only simulation and training.Conceptually, our approach is more flexible and does not need unrealistic assumptions, e.g., market clearing, that are commonly used for analytical tractability. Our GPU implementation enables training and analyzing economies with a large number of agents within reasonable time frames, e.g., training completes within a day. We demonstrate our approach in real-business-cycle models, a representative family of DGE models, with 100 worker-consumers, 10 firms, and a government who taxes and redistributes. We validate the learned meta-game ϵ-Nash equilibria through approximate best-response analyses, show that RL policies align with economic intuitions, and that our approach is constructive, e.g., by explicitly learning a spectrum of meta-game ϵ-Nash equilibria in open economic models.

****

Steerable Partial Differential Operators for Equivariant Neural Networks

Erik Jenner,Maurice Weiler

Recent work in equivariant deep learning bears strong similarities to physics. Fields over a base space are fundamental entities in both subjects, as are equivariant maps between these fields. In deep learning, however, these maps are usually defined by convolutions with a kernel, whereas they are partial differential operators (PDOs) in physics. Developing the theory of equivariant PDOs in the context of deep learning could bring these subjects even closer together and lead to a stronger flow of ideas. In this work, we derive a G-steerability constraint that completely characterizes when a PDO between feature vector fields is equivariant, for arbitrary symmetry groups G. We then fully solve this constraint for several important groups. We use our solutions as equivariant drop-in replacements for convolutional layers and benchmark them in that role. Finally, we develop a framework for equivariant maps based on Schwartz distributions that unifies classical convolutions and differential operators and gives insight about the relation between the two.

****

IntSGD: Adaptive Floatless Compression of Stochastic Gradients

Konstantin Mishchenko,Bokun Wang,Dmitry Kovalev,Peter Richtárik

We propose a family of adaptive integer compression operators for distributed Stochastic Gradient Descent (SGD) that do not communicate a single float. This is achieved by multiplying floating-point vectors with a number known to every device and then rounding to integers. In contrast to the prior work on integer compression for SwitchML by (Sapio et al., 2021), our IntSGD method is provably convergent and computationally cheaper as it estimates the scaling of vectors adaptively. Our theory shows that the iteration complexity of IntSGD matches that of SGD up to constant factors for both convex and non-convex, smooth and non-smooth functions, with and without overparameterization. Moreover, our algorithm can als o be tailored for the popular all-reduce primitive and shows promising empirical performance.

****

PAC-Bayes Information Bottleneck

Zifeng Wang,Shao-Lun Huang,Ercan Engin Kuruoglu,Jimeng Sun,Xi Chen,Yefeng Zheng

Understanding the source of the superior generalization ability of NNs remains one of the most important problems in ML research. There have been a series of theoretical works trying to derive non-vacuous bounds for NNs. Recently, the compression of information stored in weights (IIW) is proved to play a key role in NNs generalization based on the PAC-Bayes theorem. However, no solution of IIW has ever been provided, which builds a barrier for further investigation of the IIW's property and its potential in practical deep learning. In this paper, we propose an algorithm for the efficient approximation of IIW. Then, we build an IIW-based information bottleneck on the trade-off between accuracy and information complexity of NNs, namely PIB. From PIB, we can empirically identify the fitting to compressing phase transition during NNs' training and the concrete connection between the IIW compression and the generalization. Besides, we verify that IIW is able to explain NNs in broad cases, e.g., varying batch sizes, over-parameterization, and noisy labels. Moreover, we propose an MCMC-based algorithm to sample from the optimal weight posterior characterized by PIB, which fulfills the potential of IIW in enhancing NNs in practice.

****

Divergence-aware Federated Self-Supervised Learning

Weiming Zhuang,Yonggang Wen,Shuai Zhang

Self-supervised learning (SSL) is capable of learning remarkable representations from centrally available data. Recent works further implement federated learning with SSL to learn from rapidly growing decentralized unlabeled images (e.g., from cameras and phones), often resulted from privacy constraints. Extensive attention has been paid to SSL approaches based on Siamese networks. However, such an effort has not yet revealed deep insights into various fundamental building blocks for the federated self-supervised learning (FedSSL) architecture. We aim to fill in this gap via in-depth empirical study and propose a new method to tackle the non-independently and identically distributed (non-IID) data problem of decentralized data. Firstly, we introduce a generalized FedSSL framework that embraces existing SSL methods based on Siamese networks and presents flexibility catering to future methods. In this framework, a server coordinates multiple clients to conduct SSL training and periodically updates local models of clients with the aggregated global model. Using the framework, our study uncovers unique insights of FedSSL: 1) stop-gradient operation, previously reported to be essential, is not always necessary in FedSSL; 2) retaining local knowledge of clients in FedSSL is particularly beneficial for non-IID data. Inspired by the insights, we then propose a new approach for model update, Federated Divergence-aware Exponential Moving Average update (FedEMA). FedEMA updates local models of clients adaptively using EMA of the global model, where the decay rate is dynamically measured by model divergence. Extensive experiments demonstrate that FedEMA outperforms existing methods by 3-4% on linear evaluation. We hope that this work will provide useful insights for future research.

****

Classical and Quantum Algorithms for Orthogonal Neural Networks

Jonas Landman,Natansh Mathur,Iordanis Kerenidis

Orthogonal neural networks have recently been introduced as a new type of neural network imposing orthogonality on the weight matrices. They could achieve higher accuracy and avoid evanescent or explosive gradients for deep architectures. Several classical gradient descent methods have been proposed to preserve orthogonality while updating the weight matrices, but these techniques suffer from long running times and/or provide only approximate orthogonality. In this paper, we introduce a new type of neural network layer called Pyramidal Circuit, which implements an orthogonal matrix multiplication. It allows for gradient descent with perfect orthogonality with the same asymptotic running time as a standard fully connected layer. This algorithm is inspired by quantum computing and can theref ore be applied on a classical computer as well as on a near term quantum computer. It could become the building block for quantum neural networks and faster orthogonal neural networks.

****

Improving Adversarial Defense with Self-supervised Test-time Fine-tuning

Zhichao Huang,Chen Liu,Mathieu Salzmann,Sabine Süsstrunk,Tong Zhang

Although adversarial training and its variants currently constitute the most effective way to achieve robustness against adversarial attacks, their poor generalization limits their performance on the test samples. In this work, we propose to improve the generalization and robust accuracy of adversarially-trained networks via self-supervised test-time fine-tuning. To this end, we introduce a meta adversarial training method to find a good starting point for test-time fine-tuning. It incorporates the test-time fine-tuning procedure into the training phase and strengthens the correlation between the self-supervised and classification tasks. The extensive experiments on CIFAR10, STL10 and Tiny ImageNet using different self-supervised tasks show that our method consistently improves the robust accuracy under different attack strategies for both the white-box and black-box attacks.

****

Discovering Latent Network Topology in Contextualized Representations with Randomized Dynamic Programming

Yao Fu,Mirella Lapata

The discovery of large-scale discrete latent structures is crucial for understanding the fundamental generative processes of language. In this work, we use structured latent variables to study the representation space of contextualized embeddings and gain insight into the hidden topology of pretrained language models. However, existing methods are severely limited by issues of scalability and efficiency as working with large combinatorial spaces requires expensive memory consumption. We address this challenge by proposing a Randomized Dynamic Programming (RDP) algorithm for the approximate inference of structured models with DP-style exact computation (e.g., Forward-Backward). Our technique samples a subset of DP paths reducing memory complexity to as small as one percent. We use RDP to analyze the representation space of pretrained language models, discovering a large-scale latent network in a fully unsupervised way. The induced latent states not only serve as anchors marking the topology of the space (neighbors and connectivity), but also reveal linguistic properties related to syntax, morphology, and semantics. We also show that traversing this latent network yields unsupervised paraphrase generation.

****

Improving Generative Adversarial Networks via Adversarial Learning in Latent Space

Yang Li,Yichuan Mo,Liangliang Shi,Junchi Yan,Xiaolu Zhang,JUN ZHOU

Generative Adversarial Networks (GANs) have been widely studied as generative models, which map a latent distribution to the target distribution. Although many efforts have been made in terms of backbone architecture design, loss function, and training techniques, few results have been obtained on how the sampling in latent space can affect the final performance, and existing works on latent space mainly focus on controllability. 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 is not a continuous function in pixel space. From the above continuous mapping function perspective, on the other hand, two distant latent samples are also possible to be mapped into two close images. If the latent samples are mapped in aggregation into limited modes or even 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 by using the iterative fast gradient sign method (I-FGSM). We further propose new GAN training strategies to obtain better generation 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 ac hieve improvement in both quality and diversity.

****

Evaluating the Robustness of Time Series Anomaly and Intrusion Detection Methods against Adversarial Attacks

Shahroz Tariq,Simon S. Woo

Time series anomaly and intrusion detection are extensively studied in statistics, economics, and computer science. Over the years, numerous methods have been proposed for time series anomaly and intrusion detection using deep learning-based methods. Many of these methods demonstrate state-of-the-art performance on benchmark datasets, giving the false impression that these systems are robust and deployable in practical and industrial scenarios. In this paper, we demonstrate that state-of-the-art anomaly and intrusion detection methods can be easily fooled by adding adversarial perturbations to the sensor data. We use different scoring metrics such as prediction errors, anomaly, and classification scores over several public and private datasets belong to aerospace applications, automobiles, server machines, and cyber-physical systems. We evaluate state-of-the-art deep neural networks (DNNs) and graph neural networks (GNNs) methods, which claim to be robust against anomalies and intrusions, and find their performance can drop to as low as 0\% under adversarial attacks from Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD) methods. To the best of our knowledge, we are the first to demonstrate the vulnerabilities of anomaly and intrusion detection systems against adversarial attacks. Our code is available here: https://anonymous.4open.science/r/ICLR298

****

Variational Wasserstein gradient flow

Jiaojiao Fan,Amirhossein Taghvaei,Yongxin Chen

The gradient flow of a function over the space of probability densities with respect to the Wasserstein metric often exhibits nice properties and has been utilized in several machine learning applications. The standard approach to compute the Wasserstein gradient flow is the finite difference which discretizes the underlying space over a grid, and is not scalable. In this work, we propose a scalable proximal gradient type algorithm for Wasserstein gradient flow. The key of our method is a variational formulation of the objective function, which makes it possible to realize the JKO proximal map through a primal-dual optimization. This primal-dual problem can be efficiently solved by alternatively updating the parameters in the inner and outer loops. Our framework covers all the classical Wasserstein gradient flows including the heat equation and the porous medium equation. We demonstrate the performance and scalability of our algorithm with several numerical examples.

****

WHAT TO DO IF SPARSE REPRESENTATION LEARNING FAILS UNEXPECTEDLY?

Jingyi Yuan,Haoran Li,Erik Blasch,Yang Weng

Learning physical equations from data is essential for scientific discovery and engineering modeling. However, most of the existing methods rely on two rules: (1) learn a sparse representation to fit data and (2) check if the loss objective function satisfies error thresholds. This paper illustrates that such conditions are far from sufficient. Specifically, we show that sparse non-physical approximations exist with excellent fitting accuracy, but fail to adequately model the situation. To fundamentally resolve the data-fitting problem, we propose a physical neural network (PNN) utilizing “Range, Inertia, Symmetry, and Extrapolation” (RISE) constraints. RISE is based on a complete analysis for the generalizability of data properties for physical systems. The first three techniques focus on the definition of physics in space and time. The last technique of extrapolation is novel based on active learning without an inquiry, using cross-model validation. We validate the proposed PNN-RISE method via a synthetic dataset, power system dataset, and mass-damper system dataset. Numerical results show the universal capability of the PNN-RISE approach to quickly identify the hidden physical models without local optima, opening the door for the fast and highly accurate discovery of the physical laws or systems with external loads.

****

Collaborate to Defend Against Adversarial Attacks
Sen Cui,Jingfeng Zhang,Jian Liang,Masashi Sugiyama,Changshui Zhang

Adversarially robust learning methods require invariant predictions to a small neighborhood of its natural inputs, thus often encountering insufficient model capacity. Learning multiple models in an ensemble can mitigate this insufficiency, further improving both generalization and robustness. However, an ensemble still wastes the limited capacity of multiple models. To optimally utilizing the limited capacity, this paper proposes to learn a collaboration among multiple sub-models. Compared with the ensemble, the collaboration enables the possibility of correct predictions even if there exists a single correct sub-model. Besides, learning a collaboration could enable every sub-model to fit its own vulnerability area and reserve the rest of the sub-models to fit other vulnerability areas. To implement the idea, we propose a collaboration framework---CDA^2 the abbreviation for Collaborate to Defend against Adversarial Attacks. CDA^2 could effectively minimize the vulnerability overlap of all sub-models and then choose a representative sub-model to make correct predictions. Empirical experiments verify that CDA^2 outperforms various ensemble methods against black-box and white-box adversarial attacks.

****

Hyperparameter Tuning with Renyi Differential Privacy
Nicolas Papernot,Thomas Steinke

For many differentially private algorithms, such as the prominent noisy stochastic gradient descent (DP-SGD), the analysis needed to bound the privacy leakage of a single training run is well understood. However, few studies have reasoned about the privacy leakage resulting from the multiple training runs needed to fine tune the value of the training algorithm’s hyperparameters. In this work, we first illustrate how simply setting hyperparameters based on non-private training runs can leak private information. Motivated by this observation, we then provide privacy guarantees for hyperparameter search procedures within the framework of Renyi Differential Privacy. Our results improve and extend the work of Liu and Talwar (STOC 2019). Our analysis supports our previous observation that tuning hyperparameters does indeed leak private information, but we prove that, under certain assumptions, this leakage is modest, as long as each candidate training run needed to select hyperparameters is itself differentially private.

****

Short optimization paths lead to good generalization
Fusheng Liu,Haizhao Yang,Qianxiao Li

Optimization and generalization are two essential aspects of machine learning. In this paper, we propose a framework to connect optimization with generalization by analyzing the generalization error based on the length of optimization trajectory under the gradient flow algorithm after convergence. Through our approach, we show that, with a proper initialization, gradient flow converges following a short path with an explicit length estimate. Such an estimate induces a length-based generalization bound, showing that short optimization paths after convergence indicate good generalization. Our framework can be applied to broad settings. For example, we use it to obtain generalization estimates on three distinct machine learning models: underdetermined ℓ_p linear regression, kernel regression, and overparameterized two-layer ReLU neural networks.

****

Real-Time Neural Voice Camouflage
Mia Chiquier,Chengzhi Mao,Carl Vondrick

Automatic speech recognition systems have created exciting possibilities for applications, however they also enable opportunities for systematic eavesdropping.We propose a method to camouflage a person's voice from these systems without inconveniencing the conversation between people in the room. Standard adversarial attacks are not effective in real-time streaming situations because the characteristics of the signal will have changed by the time the attack is executed. We introduce predictive adversarial attacks, which achieves real-time performance by forecasting the attack vector that will be the most effective in the future. Under real-time constraints, our method jams the established speech recognition sys tem DeepSpeech 3.9x more than online projected gradient descent as measured through word error rate, and 6.6x more as measured through character error rate. We furthermore demonstrate our approach is practically effective in realistic environments with complex scene geometries.

****

One-Shot Generative Domain Adaptation

Ceyuan Yang,Yujun Shen,Zhiyi Zhang,Yinghao Xu,Jiapeng Zhu,Zhirong Wu,Bolei Zhou

This work aims at transferring a Generative Adversarial Network (GAN) pre-trained on one image domain to a new domain $\textit{referring to as few as just one target image}$. The main challenge is that, under limited supervision, it is extremely difficult to synthesize photo-realistic and highly diverse images, while acquiring representative characters of the target. Different from existing approaches that adopt the vanilla fine-tuning strategy, we import two lightweight modules to the generator and the discriminator respectively. Concretely, we introduce an $\textit{attribute adaptor}$ into the generator yet freeze its original parameters, through which it can reuse the prior knowledge to the most extent and hence maintain the synthesis quality and diversity. We then equip the well-learned discriminator backbone with an $\textit{attribute classifier}$ to ensure that the generator captures the appropriate characters from the reference. Furthermore, considering the poor diversity of the training data ($\textit{i.e.}$, as few as only one image), we propose to also constrain the diversity of the generative domain in the training process, alleviating the optimization difficulty. Our approach brings appealing results under various settings, $\textit{substantially}$ surpassing state-of-the-art alternatives, especially in terms of synthesis diversity. Noticeably, our method works well even with large domain gaps, and robustly converges $\textit{within a few minutes}$ for each experiment.

****

Neural Spectral Marked Point Processes

Shixiang Zhu,Haoyun Wang,Zheng Dong,Xiuyuan Cheng,Yao Xie

Self- and mutually-exciting point processes are popular models in machine learning and statistics for dependent discrete event data. To date, most existing models assume stationary kernels (including the classical Hawkes processes) and simple parametric models. Modern applications with complex event data require more general point process models that can incorporate contextual information of the events, called marks, besides the temporal and location information. Moreover, such applications often require non-stationary models to capture more complex spatio-temporal dependence. To tackle these challenges, a key question is to devise a versatile influence kernel in the point process model. In this paper, we introduce a novel and general neural network-based non-stationary influence kernel with high expressiveness for handling complex discrete events data while providing theoretical performance guarantees. We demonstrate the superior performance of our proposed method compared with the state-of-the-art on synthetic and real data.

****

How to Inject Backdoors with Better Consistency: Logit Anchoring on Clean Data

Zhiyuan Zhang,Lingjuan Lyu,Weiqiang Wang,Lichao Sun,Xu Sun

Since training a large-scale backdoored model from scratch requires a large training dataset, several recent attacks have considered to inject backdoors into a trained clean model without altering model behaviors on the clean data. Previous work finds that backdoors can be injected into a trained clean model with Adversarial Weight Perturbation (AWP), which means the variation of parameters are small in backdoor learning. In this work, we observe an interesting phenomenon that the variations of parameters are always AWPs when tuning the trained clean model to inject backdoors. We further provide theoretical analysis to explain this phenomenon. We are the first to formulate the behavior of maintaining accuracy on clean data as the consistency of backdoored models, which includes both global consistency and instance-wise consistency. We extensively analyze the effects of AWPs on the consistency of backdoored models. In order to achieve better consistency, we propose a novel anchoring loss to anchor or freeze the model behaviors on the clean data, with a theoretical guarantee.

****

Is Heterophily A Real Nightmare For Graph Neural Networks on Performing Node Classification?

Sitao Luan,Chenqing Hua,Qincheng Lu,Jiaqi Zhu,Mingde Zhao,Shuyuan Zhang,Xiao-Wen Chang,Doina Precup

Graph Neural Networks (GNNs) extend basic Neural Networks (NNs) by using the graph structures based on the relational inductive bias (homophily assumption). Though GNNs are believed to outperform NNs in real-world tasks, performance advantages of GNNs over graph-agnostic NNs seem not generally satisfactory. Heterophily has been considered as a main cause and numerous works have been put forward to address it. In this paper, we first show that not all cases of heterophily are harmful for GNNs with aggregation operation. Then, we propose new metrics based on a similarity matrix which considers the influence of both graph structure and input features on GNNs. The metrics demonstrate advantages over the commonly used homophily metrics in tests on synthetic graphs. From the metrics and the observations, we find that some cases of harmful heterophily can be addressed by diversification operation. By using this fact and knowledge of filterbanks, we propose the Adaptive Channel Mixing (ACM) framework to adaptively exploit aggregation, diversification and identity channels in each GNN layer, in order to address harmful heterophily. We validate the ACM-augmented baselines with 10 real-world node classification tasks. They consistently achieve significant performance gain and exceed the state-of-the-art GNNs on most of the tasks without incurring significant computational burden.

****

Improving Out-of-Distribution Robustness via Selective Augmentation

Huaxiu Yao,Yu Wang,Sai Li,Linjun Zhang,Weixin Liang,James Zou,Chelsea Finn

Machine learning algorithms typically assume that training and test examples are drawn from the same distribution. However, distribution shifts is a common problem in real-world applications and can cause models to perform dramatically worse at test time. In this paper, we specifically consider the problems of domain shifts and subpopulation shifts, where learning invariant representations by aligning domain-specific representations or balancing the risks across domains with regularizers are popular solutions. However, designing regularizers that are suitable for diverse real-world datasets is challenging. Instead, we shed new light on addressing distribution shifts by directly eliminating domain-related spurious correlations with augmentation, leading to a simple technique based on mixup, called LISA (Learning Invariant Representations via Selective Augmentation). LISA selectively interpolates samples either with the same labels but different domains or with the same domain but different labels. Empirically, we study the effectiveness of LISA on nine benchmarks ranging from subpopulation shifts to domain shifts. The results indicate that LISA consistently outperforms other state-of-the-art methods with superior invariant representations. The empirical findings are further strengthened by our theoretical analysis.

****

A Biologically Interpretable Graph Convolutional Network to Link Genetic Risk Pathways and Imaging Phenotypes of Disease

Sayan Ghosal,Qiang Chen,Giulio Pergola,Aaron L Goldman,William Ulrich,Daniel R Weinberger,Archana Venkataraman

We propose a novel end-to-end framework for whole-brain and whole-genome imaging-genetics. Our genetics network uses hierarchical graph convolution and pooling operations to embed subject-level data onto a low-dimensional latent space. The hierarchical network implicitly tracks the convergence of genetic risk across well-established biological pathways, while an attention mechanism automatically identifies the salient edges of this network at the subject level. In parallel, our imaging network projects multimodal data onto a set of latent embeddings. For interpretability, we implement a Bayesian feature selection strategy to extract the discriminative imaging biomarkers; these feature weights are optimized alongside the other model parameters. We couple the imaging and genetic embeddings with a predictor network, to ensure that the learned representations are linked to phenotype. We evaluate our framework on a schizophrenia dataset that includes two functional MRI paradigms and gene scores derived from Single Nucleotide Polymorphism data. Using repeated 10-fold cross-validation, we show that our imaging-genetics fusion achieves the better classification performance than state-of-the-art baselines. In an exploratory analysis, we further show that the biomarkers identified by our model are reproducible and closely associated with deficits in schizophrenia.

****

Differentiable Prompt Makes Pre-trained Language Models Better Few-shot Learners

Ningyu Zhang,Luoqiu Li,Xiang Chen,Shumin Deng,Zhen Bi,Chuanqi Tan,Fei Huang,Huajun Chen

Large-scale pre-trained language models have contributed significantly to natural language processing by demonstrating remarkable abilities as few-shot learners. However, their effectiveness depends mainly on scaling the model parameters and prompt design, hindering their implementation in most real-world applications. This study proposes a novel pluggable, extensible, and efficient approach named DifferentiAble pRompT (DART), which can convert small language models into better few-shot learners. The main principle behind this approach involves reformulating potential natural language processing tasks into the task of a pre-trained language model and differentially optimizing the prompt template as well as the target label with backpropagation. Furthermore, the proposed approach can be: (i) Plugged to any pre-trained language models; (ii) Extended to widespread classification tasks. A comprehensive evaluation of standard NLP tasks demonstrates that the proposed approach achieves a better few-shot performance.

****

OntoProtein: Protein Pretraining With Gene Ontology Embedding

Ningyu Zhang,Zhen Bi,Xiaozhuan Liang,Siyuan Cheng,Haosen Hong,Shumin Deng,Qiang Zhang,Jiazhang Lian,Huajun Chen

Self-supervised protein language models have proved their effectiveness in learning the proteins representations. With the increasing computational power, current protein language models pre-trained with millions of diverse sequences can advance the parameter scale from million-level to billion-level and achieve remarkable improvement. However, those prevailing approaches rarely consider incorporating knowledge graphs (KGs), which can provide rich structured knowledge facts for better protein representations. We argue that informative biology knowledge in KGs can enhance protein representation with external knowledge. In this work, we propose OntoProtein, the first general framework that makes use of structure in GO (Gene Ontology) into protein pre-training models. We construct a novel large-scale knowledge graph that consists of GO and its related proteins, and gene annotation texts or protein sequences describe all nodes in the graph. We propose novel contrastive learning with knowledge-aware negative sampling to jointly optimize the knowledge graph and protein embedding during pre-training. Experimental results show that OntoProtein can surpass state-of-the-art methods with pre-trained protein language models in TAPE benchmark and yield better performance compared with baselines in protein-protein interaction and protein function prediction.

****

Generative Adversarial Training for Neural Combinatorial Optimization Models

Liang Xin,Wen Song,Zhiguang Cao,Jie Zhang

Recent studies show that deep neural networks can be trained to learn good heuristics for various Combinatorial Optimization Problems (COPs). However, it remains a great challenge for the trained deep optimization models to generalize to distributions different from the training one. To address this issue, we propose a general framework, Generative Adversarial Neural Combinatorial Optimization (GANCO) which is equipped with another deep model to generate training instances for the optimization model, so as to improve its generalization ability. The two models are trained alternatively in an adversarial way, where the generation model is trained by reinforcement learning to find instance distributions hard for the optimization model. We apply the GANCO framework to two recent deep combinatorial optimization models, i.e., Attention Model (AM) and Policy Optimization with Multiple Optima (POMO). Extensive experiments on various problems such as Trav eling Salesman Problem, Capacitated Vehicle Routing Problem, and 0-1 Knapsack Problem show that GANCO can significantly improve the generalization ability of optimization models on various instance distributions, with little sacrifice of performance on the original training distribution.

****

Learning to Prompt for Vision-Language Models

Kaiyang Zhou,Jingkang Yang,Chen Change Loy,Ziwei Liu

Vision-language pre-training has recently emerged as a promising alternative for representation learning. It shifts from the tradition of using images and discrete labels for learning a fixed set of weights, seen as visual concepts, to aligning images and raw text for two separate encoders. Such a paradigm benefits from a broader source of supervision and allows zero-shot transfer to downstream tasks since visual concepts can be diametrically generated from natural language, known as prompt. In this paper, we identify that a major challenge of deploying such models in practice is prompt engineering. This is because designing a proper prompt, especially for context words surrounding a class name, requires domain expertise and typically takes a significant amount of time for words tuning since a slight change in wording could have a huge impact on performance. Moreover, different downstream tasks require specific designs, further hampering the efficiency of deployment. To overcome this challenge, we propose a novel approach named \emph{context optimization (CoOp)}. The main idea is to model context in prompts using continuous representations and perform end-to-end learning from data while keeping the pre-trained parameters fixed. In this way, the design of task-relevant prompts can be fully automated. Experiments on 11 datasets show that CoOp effectively turns pre-trained vision-language models into data-efficient visual learners, requiring as few as one or two shots to beat hand-crafted prompts with a decent margin and able to gain significant improvements when using more shots (e.g., at 16 shots the average gain is around 17\% with the highest reaching over 50\%). CoOp also exhibits strong robustness to distribution shift.

****

Riemannian Manifold Embeddings for Straight-Through Estimator

Jun Chen,Hanwen Chen,Jiangning Zhang,yuang Liu,Tianxin Huang,Yong Liu

Quantized Neural Networks (QNNs) aim at replacing full-precision weights $\boldsymbol{W}$ with quantized weights $\boldsymbol{\hat{W}}$, which make it possible to deploy large models to mobile and miniaturized devices easily. However, either infinite or zero gradients caused by non-differentiable quantization significantly affect the training of quantized models. In order to address this problem, most training-based quantization methods use Straight-Through Estimator (STE) to approximate gradients $\nabla_{\boldsymbol{W}}$ w.r.t. $\boldsymbol{W}$ with gradients $\nabla_{\boldsymbol{\hat{W}}}$ w.r.t. $\boldsymbol{\hat{W}}$ where the premise is that $\boldsymbol{W}$ must be clipped to $[-1,+1]$. However, the simple application of STE brings with the gradient mismatch problem, which affects the stability of the training process. In this paper, we propose to revise an approximated gradient for penetrating the quantization function with manifold learning. Specifically, by viewing the parameter space as a metric tensor in the Riemannian manifold, we introduce the Manifold Quantization (ManiQuant) via revised STE to alleviate the gradient mismatch problem. The ablation studies and experimental results demonstrate that our proposed method has a better and more stable performance with various deep neural networks on CIFAR10/100 and ImageNet datasets.

****

Permutation Compressors for Provably Faster Distributed Nonconvex Optimization

Rafa■ Szlendak,Alexander Tyurin,Peter Richtárik

In this work we study the MARINA method of Gorbunov et al (ICML, 2021) -- the current state-of-the-art distributed non-convex optimization method in terms of theoretical communication complexity. Theoretical superiority of this method can be largely attributed to two sources: a carefully engineered biased stochastic gradient estimator, which leads to a reduction in the number of communication rounds, and the reliance on
{\em independent} stochastic communication compression, which leads to a reduct

ion in the number of transmitted bits within each communication round. In this paper we i) extend the theory of MARINA to support a much wider class of potentially {\em correlated} compressors, extending the reach of the method beyond the classical independent compressors setting, ii) show that a new quantity, for which we coin the name {\em Hessian variance}, allows us to significantly refine the original analysis of MARINA without any additional assumptions, and iii) identify a special class of correlated compressors based on the idea of {\em random permutations}, for which we coin the term Perm$K$, the use of which leads to up to $O(\sqrt{n})$ (resp. $O(1 + d/\sqrt{n})$) improvement in the theoretical communication complexity of MARINA in the low Hessian variance regime when $d\geq n$ (resp. $d \leq n$), where $n$ is the number of workers and $d$ is the number of parameters describing the model we are learning. We corroborate our theoretical results with carefully engineered synthetic experiments with minimizing the average of nonconvex quadratics, and on autoencoder training with the MNIST dataset.

****

EF21 with Bells & Whistles: Practical Algorithmic Extensions of Modern Error Feedback

Ilyas Fatkhullin,Igor Sokolov,Eduard Gorbunov,Zhize Li,Peter Richtárik

First proposed by Seide et al (2014) as a heuristic, error feedback (EF) is a very popular mechanism for enforcing convergence of distributed gradient-based optimization methods enhanced with communication compression strategies based on the application of contractive compression operators. However, existing theory of EF relies on very strong assumptions (e.g., bounded gradients), and provides pessimistic convergence rates (e.g., while the best known rate for EF in the smooth nonconvex regime, and when full gradients are compressed, is $O(1/T^{2/3})$, the rate of gradient descent in the same regime is $O(1/T)$). Recently, Richt\'{a}rik et al (2021) proposed a new error feedback mechanism, EF21, based on the construction of a Markov compressor induced by a contractive compressor. EF21 removes the aforementioned theoretical deficiencies of EF and at the same time works better in practice. In this work we propose six practical extensions of EF21: partial participation, stochastic approximation, variance reduction, proximal setting, momentum and bidirectional compression. Our extensions are supported by strong convergence theory in the smooth nonconvex and also Polyak-■ojasiewicz regimes. Several of these techniques were never analyzed in conjunction with EF before, and in cases where they were (e.g., bidirectional compression), our rates are vastly superior.

****

LSP : Acceleration and Regularization of Graph Neural Networks via Locality Sensitive Pruning of Graphs

Eitan Kosman,Dotan Di Castro,Joel Oren

Graph Neural Networks (GNNs) have emerged as highly successful tools for graph-related tasks. However, real-world problems involve very large graphs, and the compute resources needed to fit GNNs to those problems grow rapidly. Moreover, the noisy nature and size of real-world graphs cause GNNs to over-fit if not regularized properly. Surprisingly, recent works show that large graphs often involve many redundant components that can be removed without compromising the performance too much. This includes node or edge removals during inference through GNNs layers or as a pre-processing step that sparsifies the input graph. This intriguing phenomenon enables the development of state-of-the-art GNNs that are both efficient and accurate. In this paper, we take a further step towards demystifying this phenomenon and propose a systematic method called Locality-Sensitive Pruning (LSP) for graph pruning based on Locality-Sensitive Hashing. We aim to sparsify a graph so that similar local environments of the original graph result in similar environments in the resulting sparsified graph, which is an essential feature for graph-related tasks. To justify the application of pruning based on local graph properties, we exemplify the advantage of applying pruning based on locality properties over other pruning strategies in various scenarios. Extensive experiments on synthetic and real-world datasets demonstrate the superiority of LSP, which removes a significant amount of edges from large graphs without compromi sing the performance, accompanied by a considerable acceleration.

****

Self Reward Design with Fine-grained Interpretability

Erico Tjoa,Cuntai Guan

Transparency and fairness issues in Deep Reinforcement Learning may stem from the black-box nature of deep neural networks used to learn its policy, value functions etc. This paper proposes a way to circumvent the issues through the bottom-up design of neural networks (NN) with detailed interpretability, where each neuron or layer has its own meaning and utility that corresponds to humanly understandable concept. With deliberate design, we show that lavaland problems can be solved using NN model with few parameters. Furthermore, we introduce the Self Reward Design (SRD), inspired by the Inverse Reward Design, so that our interpretable design can (1) solve the problem by pure design (although imperfectly) (2) be optimized via SRD (3) perform avoidance of unknown states by recognizing the inactivations of neurons aggregated as the activation in \(w_{unknown}\).

****

Two Instances of Interpretable Neural Network for Universal Approximations

Erico Tjoa,Cuntai Guan

This paper proposes two bottom-up interpretable neural network (NN) constructions for universal approximation, namely Triangularly-constructed NN (TNN) and Semi-Quantized Activation NN (SQANN). The notable properties are (1) resistance to catastrophic forgetting (2) existence of proof for arbitrarily high accuracies on training dataset (3) for an input x, users can identify specific samples of training data whose activation "fingerprints" are similar to that of x's activations. Users can also identify samples that are out of distribution.

****

Few-shot Learning via Dirichlet Tessellation Ensemble

Chunwei Ma,Ziyun Huang,Mingchen Gao,Jinhui Xu

Few-shot learning (FSL) is the process of rapid generalization from abundant base samples to inadequate novel samples. Despite extensive research in recent years, FSL is still not yet able to generate satisfactory solutions for a wide range of real-world applications. To confront this challenge, we study the FSL problem from a geometric point of view in this paper. One observation is that the widely embraced ProtoNet model is essentially a Voronoi Diagram (VD) in the feature space. We retrofit it by making use of a recent advance in computational geometry called Cluster-induced Voronoi Diagram (CIVD). Starting from the simplest nearest neighbor model, CIVD gradually incorporates cluster-to-point and then cluster-to-cluster relationships for space subdivision, which is used to improve the accuracy and robustness at multiple stages of FSL. Specifically, we use CIVD (1) to integrate parametric and nonparametric few-shot classifiers; (2) to combine feature representation and surrogate representation; (3) and to leverage feature-level, transformation-level, and geometry-level heterogeneities for a better ensemble. Our CIVD-based workflow enables us to achieve new state-of-the-art results on mini-ImageNet, CUB, and tiered-ImagenNet datasets, with ${\sim}2\%{-}5\%$ improvements upon the next best. To summarize, CIVD provides a mathematically elegant and geometrically interpretable framework that compensates for extreme data insufficiency, prevents overfitting, and allows for fast geometric ensemble for thousands of individual VD. These together make FSL stronger.

****

Deep Point Cloud Reconstruction

Jaesung Choe,ByeongIn Joung,Francois Rameau,Jaesik Park,In So Kweon

Point cloud obtained from 3D scanning is often sparse, noisy, and irregular. To cope with these issues, recent studies have been separately conducted to densify, denoise, and complete inaccurate point cloud. In this paper, we advocate that jointly solving these tasks leads to significant improvement for point cloud reconstruction. To this end, we propose a deep point cloud reconstruction network consisting of two stages: 1) a 3D sparse stacked-hourglass network as for the initial densification and denoising, 2) a refinement via transformers converting the discrete voxels into continuous 3D points. In particular, we further improve the performance of the transformers by a newly proposed module called amplified p ositional encoding. This module has been designed to differently amplify the magnitude of positional encoding vectors based on the points' distances for adaptive refinements. Extensive experiments demonstrate that our network achieves state-of-the-art performance among the recent studies in the ScanNet, ICL-NUIM, and ShapeNet datasets. Moreover, we underline the ability of our network to generalize toward real-world and unmet scenes.

****

β-Intact-VAE: Identifying and Estimating Causal Effects under Limited Overlap

Pengzhou Abel Wu,Kenji Fukumizu

As an important problem in causal inference, we discuss the identification and estimation of treatment effects (TEs) under limited overlap; that is, when subjects with certain features belong to a single treatment group. We use a latent variable to model a prognostic score which is widely used in biostatistics and sufficient for TEs; i.e., we build a generative prognostic model. We prove that the latent variable recovers a prognostic score, and the model identifies individualized treatment effects. The model is then learned as β-Intact-VAE––a new type of variational autoencoder (VAE). We derive the TE error bounds that enable representations balanced for treatment groups conditioned on individualized features. The proposed method is compared with recent methods using (semi-)synthetic datasets.

****

Promoting Saliency From Depth: Deep Unsupervised RGB-D Saliency Detection

Wei Ji,Jingjing Li,Qi Bi,chuan guo,Jie Liu,Li Cheng

Growing interests in RGB-D salient object detection (RGB-D SOD) have been witnessed in recent years, owing partly to the popularity of depth sensors and the rapid progress of deep learning techniques. Unfortunately, existing RGB-D SOD methods typically demand large quantity of training images being thoroughly annotated at pixel-level. The laborious and time-consuming manual annotation has become a real bottleneck in various practical scenarios. On the other hand, current unsupervised RGB-D SOD methods still heavily rely on handcrafted feature representations. This inspires us to propose in this paper a deep unsupervised RGB-D saliency detection approach, which requires no manual pixel-level annotation during training. It is realized by two key ingredients in our training pipeline. First, a depth-disentangled saliency update (DSU) framework is designed to automatically produce pseudo-labels with iterative follow-up refinements, which provides more trustworthy supervision signals for training the saliency network. Second, an attentive training strategy is introduced to tackle the issue of noisy pseudo-labels, by properly re-weighting to highlight the more reliable pseudo-labels. Extensive experiments demonstrate the superior efficiency and effectiveness of our approach in tackling the challenging unsupervised RGB-D SOD scenarios. Moreover, our approach can also be adapted to work in fully-supervised situation. Empirical studies show the incorporation of our approach gives rise to notably performance improvement in existing supervised RGB-D SOD models.

****

Byzantine-Robust Learning on Heterogeneous Datasets via Bucketing

Sai Praneeth Karimireddy,Lie He,Martin Jaggi

In Byzantine robust distributed or federated learning, a central server wants to train a machine learning model over data distributed across multiple workers. However, a fraction of these workers may deviate from the prescribed algorithm and send arbitrary messages. While this problem has received significant attention recently, most current defenses assume that the workers have identical data. For realistic cases when the data across workers are heterogeneous (non-iid), we design new attacks which circumvent current defenses, leading to significant loss of performance. We then propose a simple bucketing scheme that adapts existing robust algorithms to heterogeneous datasets at a negligible computational cost. We also theoretically and experimentally validate our approach, showing that combining bucketing with existing robust algorithms is effective against challenging attacks. Our work is the first to establish guaranteed convergence for the non

-iid Byzantine robust problem under realistic assumptions.

****

Distributed Methods with Compressed Communication for Solving Variational Inequalities, with Theoretical Guarantees

Aleksandr Beznosikov,Peter Richtárik,Michael Diskin,Max Ryabinin,Alexander Gasnikov

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 these and other applications, it is necessary 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 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. We empirically validate our conclusions using two experimental setups: a standard bilinear min-max problem, and large-scale distributed adversarial training of transformers.

****

Retriever: Learning Content-Style Representation as a Token-Level Bipartite Graph

Dacheng Yin,Xuanchi Ren,Chong Luo,Yuwang Wang,Zhiwei Xiong,Wenjun Zeng

This paper addresses the unsupervised learning of content-style decomposed representation. We first give a definition of style and then model the content-style representation as a token-level bipartite graph. An unsupervised framework, named Retriever, is proposed to learn such representations. First, a cross-attention module is employed to retrieve permutation invariant (P.I.) information, defined as style, from the input data. Second, a vector quantization (VQ) module is used, together with man-induced constraints, to produce interpretable content tokens. Last, an innovative link attention module serves as the decoder to reconstruct data from the decomposed content and style, with the help of the linking keys. Being modal-agnostic, the proposed Retriever is evaluated in both speech and image domains. The state-of-the-art zero-shot voice conversion performance confirms the disentangling ability of our framework. Top performance is also achieved in the part discovery task for images, verifying the interpretability of our representation. In addition, the vivid part-based style transfer quality demonstrates the potential of Retriever to support various fascinating generative tasks. Project page at https://ydcustc.github.io/retriever-demo/.

****

Neural Markov Controlled SDE: Stochastic Optimization for Continuous-Time Data

Sung Woo Park,Kyungjae Lee,Junseok Kwon

We propose a novel probabilistic framework for modeling stochastic dynamics with the rigorous use of stochastic optimal control theory. The proposed model called the neural Markov controlled stochastic differential equation (CSDE) overcomes the fundamental and structural limitations of conventional dynamical models by introducing the following two components: (1) Markov dynamic programming to efficiently train the proposed CSDE and (2) multi-conditional forward-backward losses to provide rich information for accurate inference and to assure theoretical optimality. We demonstrate that our dynamical model efficiently generates a complex time series in the data space without extra networks while showing comparable performance against existing model-based methods on several datasets.

****

Multi-modal Self-supervised Pre-training for Regulatory Genome Across Cell Types

Shentong Mo,Xi Fu,Chenyang Hong,Yizhen Chen,Yuxuan Zheng,Xiangru Tang,Yanyan Lan

,Zhiqiang Shen,Eric Xing

In the genome biology research, regulatory genome modeling is an important topic for many regulatory downstream tasks, such as promoter classification, transaction factor binding sites prediction. The core problem is to model how regulatory elements interact with each other and its variability across different cell types. However, current deep learning methods often focus on modeling genome sequences of a fixed set of cell types and do not account for the interaction between multiple regulatory elements, making them only perform well on the cell types in the training set and lack the generalizability required in biological applications. In this work, we propose a simple yet effective approach for pre-training genome data in a multi-modal and self-supervised manner, which we call $\textbf{\texttt{GeneBERT}}$. Specifically, we simultaneously take the 1d sequence of genome data and a 2d matrix of (transcription factors × regions) as the input, where three pre-training tasks are proposed to improve the robustness and generalizability of our model. We pre-train our model on the ATAC-seq dataset with 17 million genome sequences. We evaluate our GeneBERT on regulatory downstream tasks across different cell types, including promoter classification, transaction factor binding sites prediction, disease risk estimation, and splicing sites prediction. Extensive experiments demonstrate the effectiveness of multi-modal and self-supervised pre-training for large-scale regulatory genomics data.

****

CrossFormer: A Versatile Vision Transformer Hinging on Cross-scale Attention

Wenxiao Wang,Lu Yao,Long Chen,Binbin Lin,Deng Cai,Xiaofei He,Wei Liu

Transformers have made great progress in dealing with computer vision tasks. However, existing vision transformers have not yet possessed the ability of building the interactions among features of different scales, which is perceptually important to visual inputs. The reasons are two-fold: (1) Input embeddings of each layer are equal-scale, so no cross-scale feature can be extracted; (2) to lower the computational cost, some vision transformers merge adjacent embeddings inside the self-attention module, thus sacrificing small-scale (fine-grained) features of the embeddings and also disabling the cross-scale interactions. To this end, we propose Cross-scale Embedding Layer (CEL) and Long Short Distance Attention (LSDA). On the one hand, CEL blends each embedding with multiple patches of different scales, providing the self-attention module itself with cross-scale features. On the other hand, LSDA splits the self-attention module into a short-distance one and a long-distance counterpart, which not only reduces the computational burden but also keeps both small-scale and large-scale features in the embeddings. Through the above two designs, we achieve cross-scale attention. Besides, we put forward a dynamic position bias for vision transformers to make the popular relative position bias apply to variable-sized images. Hinging on the cross-scale attention module, we construct a versatile vision architecture, dubbed CrossFormer, which accommodates variable-sized inputs. Extensive experiments show that CrossFormer outperforms the other vision transformers on image classification, object detection, instance segmentation, and semantic segmentation tasks.

****

DMSANET: DUAL MULTI SCALE ATTENTION NETWORK

Abhinav Sagar

Attention mechanism of late has been quite popular in the computer vision community. A lot of work has been done to improve the performance of the network, although almost always it results in increased computational complexity. In this paper, we propose a new attention module that not only achieves the best performance but also has lesser parameters compared to most existing models. Our attention module can easily be integrated with other convolutional neural networks because of its lightweight nature. The proposed network named Dual Multi Scale Attention Network (DMSANet) is comprised of two parts: the first part is used to extract features at various scales and aggregate them, the second part uses spatial and channel attention modules in parallel to adaptively integrate local features with their global dependencies. We benchmark our network performance for Image

Classification on ImageNet dataset, Object Detection and Instance Segmentation both on MS COCO dataset.

****

AASEG: ATTENTION AWARE NETWORK FOR REAL TIME SEMANTIC SEGMENTATION

Abhinav Sagar

In this paper, we present a new network named Attention Aware Network (AASeg) for real time semantic image segmentation. Our network incorporates spatial and channel information using Spatial Attention (SA) and Channel Attention (CA) modules respectively. It also uses dense local multi-scale context information using Multi Scale Context (MSC) module. The feature maps are concatenated individually to produce the final segmentation map. We demonstrate the effectiveness of our method using a comprehensive analysis, quantitative experimental results and ablation study using Cityscapes, ADE20K and Camvid datasets. Our network performs better than most previous architectures with a 74.4% Mean IOU on Cityscapes test dataset while running at 202.7 FPS.

****

UNCERTAINTY QUANTIFICATION USING VARIATIONAL INFERENCE FOR BIOMEDICAL IMAGE SEGMENTATION

Abhinav Sagar

Deep learning motivated by convolutional neural networks has been highly successful in a range of medical imaging problems like image classification, image segmentation, image synthesis etc. However for validation and interpretability, not
only do we need the predictions made by the model but also how confident it is while making those predictions. This is important in safety critical applications
for the people to accept it. In this work, we used an encoder decoder architecture
based on variational inference techniques for segmenting brain tumour images. We evaluate our work on the publicly available BRATS dataset using Dice Similarity Coefficient (DSC) and Intersection Over Union (IOU) as the evaluation metrics. Our model is able to segment brain tumours while taking into account both aleatoric
uncertainty and epistemic uncertainty in a principled bayesian manner.

****

AA-PINN: ATTENTION AUGMENTED PHYSICS INFORMED NEURAL NETWORKS

Abhinav Sagar

Physics Informed Neural Networks has been quite successful in modelling the complex nature of fluid flow. Computational Fluid Dynamics using parallel processing algorithms on GPUs have considerably reduced the time to solve the Navier Stokes Equations. CFD based approaches uses approximates to make the modelling easy but it comes at the cost of decrease in accuracy. In this paper, we propose an attention based network architecture named AA-PINN to model PDEs behind fluid flow. We use a combination of channel and spatial attention module. We propose a novel loss function which is more robust in handling the initial as well as boundary
conditions imposed. Using evaluation metrics like RMSE, divergence and thermal kinetic energy, our network outperforms previous PINNs for modelling Navier Stokes and Burgers Equation.

****

Adversarially Robust Conformal Prediction

Asaf Gendler,Tsui-Wei Weng,Luca Daniel,Yaniv Romano

Conformal prediction is a model-agnostic tool for constructing prediction sets that are valid under the common i.i.d. assumption, which has been applied to quantify the prediction uncertainty of deep net classifiers. In this paper, we generalize this framework to the case where adversaries exist during inference time, under which the i.i.d. assumption is grossly violated. By combining conformal prediction with randomized smoothing, our proposed method forms a prediction set with finite-sample coverage guarantee that holds for any data distribution with ℓ_2-norm bounded adversarial noise, generated by any adversarial attack algo rithm. The core idea is to bound the Lipschitz constant of the non-conformity score by smoothing it with Gaussian noise and leverage this knowledge to account for the effect of the unknown adversarial perturbation. We demonstrate the necessity of our method in the adversarial setting and the validity of our theoretical guarantee on three widely used benchmark data sets: CIFAR10, CIFAR100, and ImageNet.

****

Hot-Refresh Model Upgrades with Regression-Free Compatible Training in Image Retrieval

Binjie Zhang,Yixiao Ge,Yantao Shen,Yu Li,Chun Yuan,XUYUAN XU,Yexin Wang,Ying Shan

The task of hot-refresh model upgrades of image retrieval systems plays an essential role in the industry but has never been investigated in academia before. Conventional cold-refresh model upgrades can only deploy new models after the gallery is overall backfilled, taking weeks or even months for massive data. In contrast, hot-refresh model upgrades deploy the new model immediately and then gradually improve the retrieval accuracy by backfilling the gallery on-the-fly. Compatible training has made it possible, however, the problem of model regression with negative flips poses a great challenge to the stable improvement of user experience. We argue that it is mainly due to the fact that new-to-old positive query-gallery pairs may show less similarity than new-to-new negative pairs. To solve the problem, we introduce a Regression-Alleviating Compatible Training (RACT) method to properly constrain the feature compatibility while reducing negative flips. The core is to encourage the new-to-old positive pairs to be more similar than both the new-to-old negative pairs and the new-to-new negative pairs. An efficient uncertainty-based backfilling strategy is further introduced to fasten accuracy improvements. Extensive experiments on large-scale retrieval benchmarks (e.g., Google Landmark) demonstrate that our RACT effectively alleviates the model regression for one more step towards seamless model upgrades.

****

Distribution Matching in Deep Generative Models with Kernel Transfer Operators

Zhichun Huang,Rudrasis Chakraborty,Vikas Singh

Generative models which use explicit density modeling (e.g., variational autoencoders, flow-based generative models) involve finding a mapping from a known distribution, e.g. Gaussian, to the unknown input distribution. This often requires searching over a class of non-linear functions (e.g., representable by a deep neural network). While effective in practice, the associated runtime/memory costs can increase rapidly, usually as a function of the performance desired in an application. We propose a substantially cheaper (and simpler) distribution matching strategy based on adapting known results on kernel transfer operators. We show that our formulation enables highly efficient distribution approximation and sampling, and offers surprisingly good empirical performance that compares favorably with powerful baselines, but with significant runtime savings. We show that the algorithm also performs well in small sample size settings (in brain imaging).

****

Visual Representation Learning over Latent Domains

Lucas Deecke,Timothy Hospedales,Hakan Bilen

A fundamental shortcoming of deep neural networks is their specialization to a single task and domain. While multi-domain learning enables the learning of compact models that span multiple visual domains, these rely on the presence of domain labels, in turn requiring laborious curation of datasets. This paper proposes a less explored, but highly realistic new setting called latent domain learning: learning over data from different domains, without access to domain annotations. Experiments show that this setting is challenging for standard models and existing multi-domain approaches, calling for new customized solutions: a sparse adaptation strategy is formulated which enhances performance by accounting for latent domains in data. Our method can be paired seamlessly with existing models, and benefits conceptually related tasks, e.g. empirical fairness problems and long-tailed recognition.

****

Understanding the Role of Self Attention for Efficient Speech Recognition

Kyuhong Shim,Jungwook Choi,Wonyong Sung

Self-attention (SA) is a critical component of Transformer neural networks that have succeeded in automatic speech recognition (ASR). In this paper, we analyze the role of SA in Transformer-based ASR models for not only understanding the mechanism of improved recognition accuracy but also lowering the computational complexity. We reveal that SA performs two distinct roles: phonetic and linguistic localization. Especially, we show by experiments that phonetic localization in the lower layers extracts phonologically meaningful features from speech and reduces the phonetic variance in the utterance for proper linguistic localization in the upper layers. From this understanding, we discover that attention maps can be reused as long as their localization capability is preserved. To evaluate this idea, we implement the layer-wise attention map reuse on real GPU platforms and achieve up to 1.96 times speedup in inference and 33% savings in training time with noticeably improved ASR performance for the challenging benchmark on LibriSpeech dev/test-other dataset.

****

Low Entropy Deep Networks

Chris Subia-Waud,Srinandan Dasmahapatra

The movement of data between processes and memory, not arithmetic operations, dominate the energy costs of deep learning inference calculations. This work focuses on reducing these data movement costs by reducing the number of unique weights in a network. The thinking goes that if the number of unique weights is kept small enough, then the entire network can be distributed and stored on processing elements (PEs) within accelerator designs, and the data movement costs for weight reads substantially reduced. To this end, we investigate the merits of a method, which we call Weight Fixing Networks (WFN). We design the approach to realise four model outcome objectives: i) very few unique weights, ii) low-entropy weight encodings, iii) unique weight values which are amenable to energy-saving versions of hardware multiplication, and iv) lossless task-performance. Some of these goals are conflicting. To best balance these conflicts, we combine a few novel (and some well-trodden) tricks; a novel regularisation term, (i, ii) a view of clustering cost as relative distance change (i, ii, iv), and a focus on whole-network re-use of weights (i, iii). Our Imagenet experiments demonstrate lossless compression using 56x fewer unique weights and a 1.9x lower weight-space entropy than SOTA quantisation approaches.

****

Chemical-Reaction-Aware Molecule Representation Learning

Hongwei Wang,Weijiang Li,Xiaomeng Jin,Kyunghyun Cho,Heng Ji,Jiawei Han,Martin Burke

Molecule representation learning (MRL) methods aim to embed molecules into a real vector space. However, existing SMILES-based (Simplified Molecular-Input Line-Entry System) or GNN-based (Graph Neural Networks) MRL methods either take SMILES strings as input that have difficulty in encoding molecule structure information, or over-emphasize the importance of GNN architectures but neglect their generalization ability. Here we propose using chemical reactions to assist learning molecule representation. The key idea of our approach is to preserve the equivalence of molecules with respect to chemical reactions in the embedding space, i.e., forcing the sum of reactant embeddings and the sum of product embeddings to be equal for each chemical equation. This constraint is proven effective to 1) keep the embedding space well-organized and 2) improve the generalization ability of molecule embeddings. Moreover, our model can use any GNN as the molecule encoder and is thus agnostic to GNN architectures. Experimental results demonstrate that our method achieves state-of-the-art performance in a variety of downstream tasks, e.g., reaction product prediction, molecule property prediction, reaction classification, and graph-edit-distance prediction. The code is available at https://github.com/hwwang55/MolR.

****

CycleMLP: A MLP-like Architecture for Dense Prediction

Shoufa Chen,Enze Xie,Chongjian GE,Runjian Chen,Ding Liang,Ping Luo

This paper presents a simple MLP-like architecture, CycleMLP, which is a versatile backbone for visual recognition and dense predictions. As compared to modern MLP architectures, e.g. , MLP-Mixer, ResMLP, and gMLP, whose architectures are correlated to image size and thus are infeasible in object detection and segmentation, CycleMLP has two advantages compared to modern approaches. (1) It can cope with various image sizes. (2) It achieves linear computational complexity to image size by using local windows. In contrast, previous MLPs have $O(N^2)$ computations due to fully spatial connections. We build a family of models which surpass existing MLPs and even state-of-the-art Transformer-based models, e.g. Swin Transformer, while using fewer parameters and FLOPs. We expand the MLP-like models’ applicability, making them a versatile backbone for dense prediction tasks. CycleMLP achieves competitive results on object detection, instance segmentation, and semantic segmentation. In particular, CycleMLP-Tiny outperforms Swin-Tiny by 1.3% mIoU on ADE20K dataset with fewer FLOPs. Moreover, CycleMLP also shows excellent zero-shot robustness on ImageNet-C dataset.

****

Skill-based Meta-Reinforcement Learning

Taewook Nam,Shao-Hua Sun,Karl Pertsch,Sung Ju Hwang,Joseph J Lim

While deep reinforcement learning methods have shown impressive results in robot learning, their sample inefficiency makes the learning of complex, long-horizon behaviors with real robot systems infeasible. To mitigate this issue, meta-reinforcement learning methods aim to enable fast learning on novel tasks by learning how to learn. Yet, the application has been limited to short-horizon tasks with dense rewards. To enable learning long-horizon behaviors, recent works have explored leveraging prior experience in the form of offline datasets without reward or task annotations. While these approaches yield improved sample efficiency, millions of interactions with environments are still required to solve complex tasks. In this work, we devise a method that enables meta-learning on long-horizon, sparse-reward tasks, allowing us to solve unseen target tasks with orders of magnitude fewer environment interactions. Our core idea is to leverage prior experience extracted from offline datasets during meta-learning. Specifically, we propose to (1) extract reusable skills and a skill prior from offline datasets, (2) meta-train a high-level policy that learns to efficiently compose learned skills into long-horizon behaviors, and (3) rapidly adapt the meta-trained policy to solve an unseen target task. Experimental results on continuous control tasks in navigation and manipulation demonstrate that the proposed method can efficiently solve long-horizon novel target tasks by combining the strengths of meta-learning and the usage of offline datasets, while prior approaches in RL, meta-RL, and multi-task RL require substantially more environment interactions to solve the tasks.

****

Analytic-DPM: an Analytic Estimate of the Optimal Reverse Variance in Diffusion Probabilistic Models

Fan Bao,Chongxuan Li,Jun Zhu,Bo Zhang

Diffusion probabilistic models (DPMs) represent a class of powerful generative models. Despite their success, the inference of DPMs is expensive since it generally needs to iterate over thousands of timesteps. A key problem in the inference is to estimate the variance in each timestep of the reverse process. In this work, we present a surprising result that both the optimal reverse variance and the corresponding optimal KL divergence of a DPM have analytic forms w.r.t. its score function. Building upon it, we propose \textit{Analytic-DPM}, a training-free inference framework that estimates the analytic forms of the variance and KL divergence using the Monte Carlo method and a pretrained score-based model. Further, to correct the potential bias caused by the score-based model, we derive both lower and upper bounds of the optimal variance and clip the estimate for a better result. Empirically, our analytic-DPM improves the log-likelihood of various DPMs, produces high-quality samples, and meanwhile enjoys a $20\times$ to $80\times$ speed up.

****

InfinityGAN: Towards Infinite-Pixel Image Synthesis

Chieh Hubert Lin,Hsin-Ying Lee,Yen-Chi Cheng,Sergey Tulyakov,Ming-Hsuan Yang

We present InfinityGAN, a method to generate arbitrary-sized images. The problem is associated with several key challenges. First, scaling existing models to an arbitrarily large image size is resource-constrained, both in terms of computation and availability of large-field-of-view training data. InfinityGAN trains and infers patch-by-patch seamlessly with low computational resources. Second, large images should be locally and globally consistent, avoid repetitive patterns, and look realistic. To address these, InfinityGAN takes global appearance, local structure and texture into account. With this formulation, we can generate images with spatial size and level of detail not attainable before. Experimental evaluation supports that InfinityGAN generates images with superior global structure compared to baselines and features parallelizable inference. Finally, we show several applications unlocked by our approach, such as fusing styles spatially, multi-modal outpainting and image inbetweening at arbitrary input and output sizes.

****

Flashlight: Enabling Innovation in Tools for Machine Learning

Jacob Kahn,Vineel Pratap,Tatiana Likhomanenko,Qiantong Xu,Awni Hannun,Jeff Cai,Paden Tomasello,Ann Lee,Edouard Grave,Gilad Avidov,Benoit Steiner,Vitaliy Liptchinsky,Gabriel Synnaeve,Ronan Collobert

As the computational requirements for machine learning systems and the size and complexity of machine learning frameworks increases, essential framework innovation has become challenging. While computational needs have driven recent compiler, networking, and hardware advancements, utilization of those advancements by machine learning tools is occurring at a slower pace. This is in part due to the difficulties involved in prototyping new computational paradigms with existing frameworks. Large frameworks prioritize machine learning researchers and practitioners as end users and pay comparatively little attention to systems researchers who can push frameworks forward --- we argue that both are equally-important stakeholders. We introduce Flashlight, an open source library built to spur innovation in machine learning tools and systems by prioritizing open, modular, customizable internals and state-of-the-art, research-ready models and training setups across a variety of domains. Flashlight enables systems researchers to rapidly prototype and experiment with novel ideas in machine learning computation and has low overhead, competing with and often outperforming other popular machine learning frameworks. We see Flashlight as a tool enabling research that can benefit widely-used libraries downstream and bring machine learning and systems researchers closer together.

****

Label Encoding for Regression Networks

Deval Shah,Zi Yu Xue,Tor Aamodt

Deep neural networks are used for a wide range of regression problems. However, there exists a significant gap in accuracy between specialized approaches and generic direct regression in which a network is trained by minimizing the squared or absolute error of output labels. Prior work has shown that solving a regression problem with a set of binary classifiers can improve accuracy by utilizing well-studied binary classification algorithms. We introduce binary-encoded labels (BEL), which generalizes the application of binary classification to regression by providing a framework for considering arbitrary multi-bit values when encoding target values. We identify desirable properties of suitable encoding and decoding functions used for the conversion between real-valued and binary-encoded labels based on theoretical and empirical study. These properties highlight a tradeoff between classification error probability and error-correction capabilities of label encodings. BEL can be combined with off-the-shelf task-specific feature extractors and trained end-to-end. We propose a series of sample encoding, decoding, and training loss functions for BEL and demonstrate they result in lower error than direct regression and specialized approaches while being suitable for a diverse set of regression problems, network architectures, and evaluation metri cs. BEL achieves state-of-the-art accuracies for several regression benchmarks. Code is available at https://github.com/ubc-aamodt-group/BEL_regression.

****

Exploring the Robustness of Distributional Reinforcement Learning against Noisy State Observations

Ke Sun,Yi Liu,Yingnan Zhao,Hengshuai Yao,SHANGLING JUI,Linglong Kong

In real scenarios, state observations that an agent observes may contain measurement errors or adversarial noises, misleading the agent to take suboptimal actions or even collapse while training. In this paper, we study the training robustness of distributional Reinforcement Learning~(RL), a class of state-of-the-art methods that estimate the whole distribution, as opposed to only the expectation, of the total return. Firstly, we propose State-Noisy Markov Decision Process~(SN-MDP) in the tabular case to incorporate both random and adversarial state observation noises, in which the contraction of both expectation-based and distributional Bellman operators is derived. Beyond SN-MDP with the function approximation, we theoretically characterize the bounded gradient norm of histogram-based distributional loss, accounting for the better training robustness of distribution RL. We also provide stricter convergence conditions of the Temporal-Difference~(TD) learning under more flexible state noises, as well as the sensitivity analysis by the leverage of influence function. Finally, extensive experiments on the suite of games show that distributional RL enjoys better training robustness compared with its expectation-based counterpart across various state observation noises.

****

Shuffle Private Stochastic Convex Optimization

Albert Cheu,Matthew Joseph,Jieming Mao,Binghui Peng

In shuffle privacy, each user sends a collection of randomized messages to a trusted shuffler, the shuffler randomly permutes these messages, and the resulting shuffled collection of messages must satisfy differential privacy. Prior work in this model has largely focused on protocols that use a single round of communication to compute algorithmic primitives like means, histograms, and counts. In this work, we present interactive shuffle protocols for stochastic convex optimization. Our optimization protocols rely on a new noninteractive protocol for summing vectors of bounded ℓ_2 norm. By combining this sum subroutine with techniques including mini-batch stochastic gradient descent, accelerated gradient descent, and Nesterov's smoothing method, we obtain loss guarantees for a variety of convex loss functions that significantly improve on those of the local model and sometimes match those of the central model.

****

RISP: Rendering-Invariant State Predictor with Differentiable Simulation and Rendering for Cross-Domain Parameter Estimation

Pingchuan Ma,Tao Du,Joshua B. Tenenbaum,Wojciech Matusik,Chuang Gan

This work considers identifying parameters characterizing a physical system's dynamic motion directly from a video whose rendering configurations are inaccessible. Existing solutions require massive training data or lack generalizability to unknown rendering configurations. We propose a novel approach that marries domain randomization and differentiable rendering gradients to address this problem. Our core idea is to train a rendering-invariant state-prediction (RISP) network that transforms image differences into state differences independent of rendering configurations, e.g., lighting, shadows, or material reflectance. To train this predictor, we formulate a new loss on rendering variances using gradients from differentiable rendering. Moreover, we present an efficient, second-order method to compute the gradients of this loss, allowing it to be integrated seamlessly into modern deep learning frameworks. We evaluate our method in rigid-body and deformable-body simulation environments using four tasks: state estimation, system identification, imitation learning, and visuomotor control. We further demonstrate the efficacy of our approach on a real-world example: inferring the state and action sequences of a quadrotor from a video of its motion sequences. Compared with existing methods, our approach achieves significantly lower reconstruct ion errors and has better generalizability among unknown rendering configurations.

****

Online MAP Inference and Learning for Nonsymmetric Determinantal Point Processes

Aravind Reddy,Ryan Rossi,Zhao Song,Anup Rao,Tung Mai,Nedim Lipka,Gang Wu,Eunyee Koh,Nesreen Ahmed

In this paper, we introduce the online and streaming MAP inference and learning problems for Non-symmetric Determinantal Point Processes (NDPPs) where data points arrive in an arbitrary order and the algorithms are constrained to use a single-pass over the data as well as sub-linear memory. The online setting has an additional requirement of maintaining a valid solution at any point in time. For solving these new problems, we propose algorithms with theoretical guarantees, evaluate them on several real-world datasets, and show that they give comparable performance to state-of-the-art offline algorithms that store the entire data in memory and take multiple passes over it.

****

Know Your Action Set: Learning Action Relations for Reinforcement Learning

Ayush Jain,Norio Kosaka,Kyung-Min Kim,Joseph J Lim

Intelligent agents can solve tasks in various ways depending on their available set of actions. However, conventional reinforcement learning (RL) assumes a fixed action set. This work asserts that tasks with varying action sets require reasoning of the relations between the available actions. For instance, taking a nail-action in a repair task is meaningful only if a hammer-action is also available. To learn and utilize such action relations, we propose a novel policy architecture consisting of a graph attention network over the available actions. We show that our model makes informed action decisions by correctly attending to other related actions in both value-based and policy-based RL. Consequently, it outperforms non-relational architectures on applications where the action space often varies, such as recommender systems and physical reasoning with tools and skills. Results and code at https://sites.google.com/view/varyingaction .

****

Equivariant Transformers for Neural Network based Molecular Potentials

Philipp Thölke,Gianni De Fabritiis

The prediction of quantum mechanical properties is historically plagued by a trade-off between accuracy and speed. Machine learning potentials have previously shown great success in this domain, reaching increasingly better accuracy while maintaining computational efficiency comparable with classical force fields. In this work we propose TorchMD-NET, a novel equivariant Transformer (ET) architecture, outperforming state-of-the-art on MD17, ANI-1, and many QM9 targets in both accuracy and computational efficiency. Through an extensive attention weight analysis, we gain valuable insights into the black box predictor and show differences in the learned representation of conformers versus conformations sampled from molecular dynamics or normal modes. Furthermore, we highlight the importance of datasets including off-equilibrium conformations for the evaluation of molecular potentials.

****

FROB: Few-shot ROBust Model for Classification with Out-of-Distribution Detection

Nikolaos Dionelis,Mehrdad Yaghoobi,Sotirios A. Tsaftaris

Nowadays, classification and Out-of-Distribution (OoD) detection in the few-shot setting remain challenging aims mainly due to rarity and the limited samples in the few-shot setting, and because of adversarial attacks. Accomplishing these aims is important for critical systems in safety, security, and defence. In parallel, OoD detection is challenging since deep neural network classifiers set high confidence to OoD samples away from the training data. To address such limitations, we propose the Few-shot ROBust (FROB) model for classification and few-shot OoD detection. We devise a methodology for improved robustness and reliable confidence prediction for few-shot OoD detection. We generate the support boundary of the normal class distribution and combine it with few-shot Outlier Exposure (OE). We propose a self-supervised learning few-shot confidence boundary methodol ogy based on generative and discriminative models, including classification. The main contribution of FROB is the combination of the generated boundary in a self-supervised learning manner and the imposition of low confidence at this learned boundary. FROB implicitly generates strong adversarial samples on the boundary and forces samples from OoD, including our boundary, to be less confident by the classifier. FROB achieves generalization to unseen anomalies and adversarial attacks, with applicability to unknown, in the wild, test sets that do not correlate to the training datasets. To improve robustness, FROB redesigns and streamlines OE to work even for zero-shots. By including our learned boundary, FROB effectively reduces the threshold linked to the model’s few-shot robustness, and maintains the OoD performance approximately constant and independent of the number of few-shot samples. The few-shot robustness analysis evaluation of FROB on different image sets and on One-Class Classification (OCC) data shows that FROB achieves competitive state-of-the-art performance and outperforms benchmarks in terms of robustness to the outlier OoD few-shot sample population and variability.

****

Learning to Collaborate

Sen Cui,Jian Liang,Weishen Pan,Kun Chen,Changshui Zhang,Fei Wang

In this paper, we focus on effective learning over a collaborative research network involving multiple clients. Each client has its own sample population which may not be shared with other clients due to privacy concerns. The goal is to learn a model for each client, which behaves better than the one learned from its own data, through secure collaborations with other clients in the network. Due to the discrepancies of the sample distributions across different clients, it is not necessarily that collaborating with everyone will lead to the best local models. We propose a learning to collaborate framework, where each client can choose to collaborate with certain members in the network to achieve a ``collaboration equilibrium", where smaller collaboration coalitions are formed within the network so that each client can obtain the model with the best utility. We propose the concept of benefit graph which describes how each client can benefit from collaborating with other clients and develop a Pareto optimization approach to obtain it. Finally the collaboration coalitions can be derived from it based on graph operations. Our framework provides a new way of setting up collaborations in a research network. Experiments on both synthetic and real world data sets are provided to demonstrate the effectiveness of our method.

****

Transfer Learning for Bayesian HPO with End-to-End Meta-Features

Hadi Samer Jomaa,Sebastian Pineda Arango,Lars Schmidt-Thieme,Josif Grabocka

Hyperparameter optimization (HPO) is a crucial component of deploying machine learning models, however, it remains an open problem due to the resource-constrained number of possible hyperparameter evaluations. As a result, prior work focus on exploring the direction of transfer learning for tackling the sample inefficiency of HPO. In contrast to existing approaches, we propose a novel Deep Kernel Gaussian Process surrogate with Landmark Meta-features (DKLM) that can be jointly meta-trained on a set of source tasks and then transferred efficiently on a new (unseen) target task. We design DKLM to capture the similarity between hyperparameter configurations with an end-to-end meta-feature network that embeds the set of evaluated configurations and their respective performance. As a result, our novel DKLM can learn contextualized dataset-specific similarity representations for hyperparameter configurations. We experimentally validate the performance of DKLM in a wide range of HPO meta-datasets from OpenML and demonstrate the empirical superiority of our method against a series of state-of-the-art baselines.

****

Meta-OLE: Meta-learned Orthogonal Low-Rank Embedding

Ze Wang,Yue Lu,Qiang Qiu

We introduce Meta-OLE, a new geometry-regularized method for fast adaptation to novel tasks in few-shot image classification. The proposed method learns to adapt for each few-shot classification task a feature space with simultaneous inter-class orthogonality and intra-class low-rankness. Specifically, a deep feature extractor is trained by explicitly imposing orthogonal low-rank subspace structur es among features corresponding to different classes within a given task. To adapt to novel tasks with unseen categories, we further meta-learn a light-weight transformation to enhance the inter-class margins. As an additional benefit, this light-weight transformation lets us exploit the query data for label propagation from labeled to unlabeled data without any auxiliary network components. The explicitly geometry-regularized feature subspaces allow the classifiers on novel tasks to be inferred in a closed form, with an adaptive subspace truncation that selectively discards non-discriminative dimensions. We perform experiments on standard few-shot image classification tasks, and observe performance superior to state-of-the-art meta-learning methods.

****

Fast Adaptive Anomaly Detection

Ze Wang,Yipin Zhou,Rui Wang,Tsung-Yu Lin,Ashish Shah,Ser-Nam Lim

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 workin 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 associates low energies to correct values and higher energies to incorrect values. As its core, the EBM em-ploys 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.

****

On the Importance of Difficulty Calibration in Membership Inference Attacks

Lauren Watson,Chuan Guo,Graham Cormode,Alexandre Sablayrolles

The vulnerability of machine learning models to membership inference attacks has received much attention in recent years. However, existing attacks mostly remain impractical due to having high false positive rates, where non-member samples are often erroneously predicted as members. This type of error makes the predicted membership signal unreliable, especially since most samples are non-members in real world applications. In this work, we argue that membership inference attacks can benefit drastically from difficulty calibration, where an attack's predicted membership score is adjusted to the difficulty of correctly classifying the target sample. We show that difficulty calibration can significantly reduce the false positive rate of a variety of existing attacks without a loss in accuracy.

****

Entroformer: A Transformer-based Entropy Model for Learned Image Compression

Yichen Qian,Xiuyu Sun,Ming Lin,Zhiyu Tan,Rong Jin

One critical component in lossy deep image compression is the entropy model, which predicts the probability distribution of the quantized latent representation in the encoding and decoding modules. Previous works build entropy models upon convolutional neural networks which are inefficient in capturing global dependencies. In this work, we propose a novel transformer-based entropy model, termed Entroformer, to capture long-range dependencies in probability distribution estimation effectively and efficiently. Different from vision transformers in image classification, the Entroformer is highly optimized for image compression, includi ng a top-k self-attention and a diamond relative position encoding. Meanwhile, we further expand this architecture with a parallel bidirectional context model to speed up the decoding process. The experiments show that the Entroformer achieves state-of-the-art performance on image compression while being time-efficient.

****

Accelerating Training of Deep Spiking Neural Networks with Parameter Initialization

Jianhao Ding,Jiyuan Zhang,Zhaofei Yu,Tiejun Huang

Despite that spiking neural networks (SNNs) show strong advantages in information encoding, power consuming, and computational capability, the underdevelopment of supervised learning algorithms is still a hindrance for training SNN. Our consideration is that proper weight initialization is a pivotal issue for efficient SNN training. It greatly influences gradient generating with the method of back-propagation through time at the initial training stage. Focusing on the properties of spiking neurons, we first derive the asymptotic formula of their response curve approximating the actual neuron response distribution. Then, we propose an initialization method obtained from the slant asymptote to overcome gradient vanishing. Finally, experiments with different coding schemes on classification tasks show that our method can effectively improve training speed and the final model accuracy compared with traditional deep learning initialization methods and existing SNN initialization methods. Further validation on different neuron types and training hyper-parameters has shown comparably good versatility and superiority over the other methods. Some suggestions are given to SNN training based on the analyses.

****

Communicate Then Adapt: An Effective Decentralized Adaptive Method for Deep Training

Bicheng Ying,Kun Yuan,Yiming Chen,Hanbin Hu,Yingya Zhang,Pan Pan,Wotao Yin

Decentralized adaptive gradient methods, in which each node averages only with its neighbors, are critical to save communication and wall-clock training time in deep learning tasks. While different in concrete recursions, existing decentralized adaptive methods share the same algorithm structure: each node scales its gradient with information of the past squared gradients (which is referred to as the adaptive step) before or while it communicates with neighbors. In this paper, we identify the limitation of such adapt-then/while-communicate structure: it will make the developed algorithms highly sensitive to heterogeneous data distributions, and hence deviate their limiting points from the stationary solution. To overcome this limitation, we propose an effective decentralized adaptive method with a communicate-then-adapt structure, in which each node conducts the adaptive step after finishing the neighborhood communications. The new method is theoretically guaranteed to approach to the stationary solution in the non-convex scenario. Experimental results on a variety of CV/NLP tasks show that our method has a clear superiority to other existing decentralized adaptive methods.

****

RoMA: a Method for Neural Network Robustness Measurement and Assessment

Natan Levy,Guy Katz

Neural network models have become the leading solution for various tasks, such as classification, language processing, protein folding, and others. However, their reliability is heavily plagued by adversarial inputs: small input perturbations that cause the model to produce erroneous output, thus impairing the model’s robustness.

Adversarial inputs can occur naturally when the system’s environment behaves randomly, even in the absence of a malicious adversary, and are thus a severe cause for concern when attempting to deploy neural networks within critical systems. In this paper, we present a new statistical method, called Robustness Measurement and Assessment (RoMA), which can accurately measure the robustness of a neural network model. Specifically, RoMA determines the probability that a random input perturbation might cause misclassification. The method allows us to provide formal guarantees regarding the expected number of errors a trained model will have after deployment. Our approach can be implemented on large-scale, black-box neural networks, which is a significant advantage compared to recently proposed verification methods. We apply our approach in two ways: comparing the robustness of different models, and measuring how a model’s robustness is affected by the scale of adversarial perturbation. One interesting insight obtained through this work is that, in a classification network, different output labels can exhibit very different robustness levels. We term this phenomenon Categorial Robustness. Our ability to perform risk and robustness assessments on a categorial basis opens the door to risk mitigation, which may prove to be a significant step towards neural network certification in safety-critical applications.

****

Dual Lottery Ticket Hypothesis

Yue Bai,Huan Wang,ZHIQIANG TAO,Kunpeng Li,Yun Fu

Fully exploiting the learning capacity of neural networks requires overparameterized dense networks. On the other side, directly training sparse neural networks typically results in unsatisfactory performance. Lottery Ticket Hypothesis (LTH) provides a novel view to investigate sparse network training and maintain its capacity. Concretely, it claims there exist winning tickets from a randomly initialized network found by iterative magnitude pruning and preserving promising trainability (or we say being in trainable condition). In this work, we regard the winning ticket from LTH as the subnetwork which is in trainable condition and its performance as our benchmark, then go from a complementary direction to articulate the Dual Lottery Ticket Hypothesis (DLTH): Randomly selected subnetworks from a randomly initialized dense network can be transformed into a trainable condition and achieve admirable performance compared with LTH --- random tickets in a given lottery pool can be transformed into winning tickets. Specifically, by using uniform-randomly selected subnetworks to represent the general cases, we propose a simple sparse network training strategy, Random Sparse Network Transformation (RST), to substantiate our DLTH. Concretely, we introduce a regularization term to borrow learning capacity and realize information extrusion from the weights which will be masked. After finishing the transformation for the randomly selected subnetworks, we conduct the regular finetuning to evaluate the model using fair comparisons with LTH and other strong baselines. Extensive experiments on several public datasets and comparisons with competitive approaches validate our DLTH as well as the effectiveness of the proposed model RST. Our work is expected to pave a way for inspiring new research directions of sparse network training in the future. Our code is available at https://github.com/yueb17/DLTH.

****

Collaboration of Experts: Achieving 80% Top-1 Accuracy on ImageNet with 100M FLOPs

Yikang Zhang,Zhuo Chen,Zhao Zhong

In this paper, we propose a Collaboration of Experts (CoE) framework to pool together the expertise of multiple networks towards a common aim. Each expert is an individual network with expertise on a unique portion of the dataset, which enhances the collective capacity. Given a sample, an expert is selected by the delegator, which simultaneously outputs a rough prediction to support early termination. To make each model in CoE play its role, we propose a novel training algorithm that consists of three components: weight generation module (WGM), label generation module (LGM) and selection reweighting module (SRM). Our method achieves the state-of-the-art performance on ImageNet, 80.7% top-1 accuracy with 194M FLOPs. Combined with PWLU activation function and CondConv, CoE further achieves the accuracy of 80.0% with only 100M FLOPs for the first time. More importantly, CoE is hardware-friendly, achieving a 3~6x speedup compared with some existing c onditional computation approaches. Experimental results on translation task also show the strong generalizability of CoE.

****

Decouple and Reconstruct: Mining Discriminative Features for Cross-domain Object Detection

Jiawei Wang,Konghuai Shen,Shao Ming,Jun Yin,Ming Liu

In recent years, a great progress has been witnessed for cross-domain object detection. Most state-of-the-art methods strive to handle the relation between local regions by calibrating cross-channel and spatial information to enable better alignment. They succeed in improving the generalization of the model, but implicitly drive networks to pay more attention on the shared attributes and ignore the domain-specific feature, which limits the performance of the algorithm. In order to search for the equilibrium between transferability and discriminability, we propose a novel adaptation framework for cross-domain object detection. Specifically, we adopt a style-aware feature fusion method and design two plug-and-play feature component regularization modules, which repositions the focus of the model on domain-specific features by restructuring the style and content of features. Our key insight is that while it is difficult to extract discriminative features in target domain, it is feasible to assign the underlying details to the model via feature style transfer. Without bells and whistles, our method significantly boosts the performance of existing Domain Adaptive Faster R-CNN detectors, and achieves state-of-the-art results on several benchmark datasets for cross-domain object detection.

****

Understanding the Success of Knowledge Distillation -- A Data Augmentation Perspective

Huan Wang,Suhas Lohit,Michael Jeffrey Jones,Yun Fu

Knowledge distillation (KD) is a general neural network training approach that uses a teacher model to guide a student model. Many works have explored the rationale for its success. However, its interplay with data augmentation (DA) has not been well understood so far. In this paper, we are motivated by an interesting observation in classification: KD loss can take more advantage of a DA method than cross-entropy loss \emph{simply by training for more iterations}. We present a generic framework to explain this interplay between KD and DA. Inspired by it, we enhance KD via stronger data augmentation schemes named TLmixup and TLCutMix. Furthermore, an even stronger and efficient DA approach is developed specifically for KD based on the idea of active learning. The findings and merits of our method are validated with extensive experiments on CIFAR-100, Tiny ImageNet, and ImageNet datasets. We achieve new state-of-the-art accuracy by using the original KD loss armed with stronger augmentation schemes, compared to existing state-of-the-art methods that employ more advanced distillation losses. We also show that, by combining our approaches with the advanced distillation losses, we can advance the state-of-the-art even further. In addition to very promising performance, this paper importantly sheds light on explaining the success of knowledge distillation. The interaction of KD and DA methods we have discovered can inspire more powerful KD algorithms.

****

Structured Pruning Meets Orthogonality

Huan Wang,Yun Fu

Several recent works empirically found finetuning learning rate is crucial to the final performance in structured neural network pruning. It is shown that the \emph{dynamical isometry} broken by pruning answers for this phenomenon. How to develop a filter pruning method that maintains or recovers dynamical isometry \emph{and} is scalable to modern deep networks remains elusive up to now. In this paper, we present \emph{orthogonality preserving pruning} (OPP), a regularization-based structured pruning method that maintains the dynamical isometry during pruning. Specifically, OPP regularizes the gram matrix of convolutional kernels to encourage kernel orthogonality among the important filters meanwhile driving the unimportant weights towards zero. We also propose to regularize batch-normalization parameters for better preserving dynamical isometry for the whole network.

Empirically, OPP can compete with the \emph{ideal} dynamical isometry recovery method on linear networks. On non-linear networks (ResNet56/VGG19, CIFAR datasets), it outperforms the available solutions \emph{by a large margin}. Moreover, OPP can also work effectively with modern deep networks (ResNets) on ImageNet, delivering encouraging performance in comparison to many recent filter pruning methods. To our best knowledge, this is the \emph{first} method that effectively maintains dynamical isometry during pruning for \emph{large-scale} deep neural networks.

****

Rethinking Again the Value of Network Pruning -- A Dynamical Isometry Perspective

Huan Wang,Can Qin,Yue Bai,Yun Fu

Several recent works questioned the value of inheriting weight in structured neural network pruning because they empirically found training from scratch can match or even outperform finetuning a pruned model. In this paper, we present evidences that this argument is actually \emph{inaccurate} because of using improperly small finetuning learning rates. With larger learning rates, our results consistently suggest pruning outperforms training from scratch on multiple networks (ResNets, VGG11) and datasets (MNIST, CIFAR10, ImageNet) over most pruning ratios. To deeply understand why finetuning learning rate holds such a critical role, we examine the theoretical reason behind through the lens of \emph{dynamical isometry}, a nice property of networks that can make the gradient signals preserve norm during propagation. Our results suggest that weight removal in pruning breaks dynamical isometry, \emph{which fundamentally answers for the performance gap between a large finetuning LR and~a small one}. Therefore, it is necessary to recover the dynamical isometry before finetuning. In this regard, we also present a regularization-based technique to do so, which is rather simple-to-implement yet effective in dynamical isometry recovery on modern residual convolutional neural networks.

****

Non-deep Networks

Ankit Goyal,Alexey Bochkovskiy,Jia Deng,Vladlen Koltun

Depth is the hallmark of deep neural networks. But more depth means more sequential computation and higher latency. 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.

****

SGD Can Converge to Local Maxima

Liu Ziyin,Botao Li,James B Simon,Masahito Ueda

Previous works on stochastic gradient descent (SGD) often focus on its success. In this work, we construct worst-case optimization problems illustrating that, when not in the regimes that the previous works often assume, SGD can exhibit many strange and potentially undesirable behaviors. Specifically, we construct landscapes and data distributions such that (1) SGD converges to local maxima, (2) SGD escapes saddle points arbitrarily slowly, (3) SGD prefers sharp minima over flat ones, and (4) AMSGrad converges to local maxima. We also realize results in a minimal neural network-like example. Our results highlight the importance of simultaneously analyzing the minibatch sampling, discrete-time updates rules, and realistic landscapes to understand the role of SGD in deep learning.

****

On Invariance Penalties for Risk Minimization

Kia Khezeli,Arno Blaas,Frank Soboczenski,Nicholas Chia,John Kalantari

The Invariant Risk Minimization (IRM) principle was first proposed by Arjovsky et al. (2019) to address the domain generalization problem by leveraging data heterogeneity from differing experimental conditions. Specifically, IRM seeks to find a data representation under which an optimal classifier remains invariant across all domains. Despite the conceptual appeal of IRM, the effectiveness of the originally proposed invariance penalty has recently been brought into question through stylized experiments and counterexamples. In this work, we investigate the relationship between the data representation, invariance penalty, and risk. In doing so, we propose a novel invariance penalty, and utilize it to design an adaptive rule for tuning the coefficient of the penalty proposed by Arjovsky et al. (2019). More- over, we provide practical insights on how to avoid the potential failure of IRM considered in the nascent counterexamples. Finally, we conduct numerical experiments on both synthetic and real-world data sets with the objective of building invariant predictors. In our non-synthetic experiments, we sought to build a predictor of human health status using a collection of data sets from various studies which investigate the relationship between human gut microbiome and a particular disease. We substantiate the effectiveness of our proposed approach on these data sets and thus further facilitate the adoption of the IRM principle in other real-world applications.

****

GNN is a Counter? Revisiting GNN for Question Answering

Kuan Wang,Yuyu Zhang,Diyi Yang,Le Song,Tao Qin

Question Answering (QA) has been a long-standing research topic in AI and NLP fields, and a wealth of studies has been conducted to attempt to equip QA systems with human-level reasoning capability. To approximate the complicated human reasoning process, state-of-the-art QA systems commonly use pre-trained language models (LMs) to access knowledge encoded in LMs together with elaborately designed modules based on Graph Neural Networks (GNNs) to perform reasoning over knowledge graphs (KGs). However, many problems remain open regarding the reasoning functionality of these GNN-based modules. Can these GNN-based modules really perform a complex reasoning process? Are they under- or over-complicated for QA? To open the black box of GNN and investigate these problems, we dissect state-of-the-art GNN modules for QA and analyze their reasoning capability. We discover that even a very simple graph neural counter can outperform all the existing GNN modules on CommonsenseQA and OpenBookQA, two popular QA benchmark datasets which heavily rely on knowledge-aware reasoning. Our work reveals that existing knowledge-aware GNN modules may only carry out some simple reasoning such as counting. It remains a challenging open problem to build comprehensive reasoning modules for knowledge-powered QA.

****

IFR-Explore: Learning Inter-object Functional Relationships in 3D Indoor Scenes

QI LI,Kaichun Mo,Yanchao Yang,Hang Zhao,Leonidas Guibas

Building embodied intelligent agents that can interact with 3D indoor environments has received increasing research attention in recent years. While most works focus on single-object or agent-object visual functionality and affordances, our work proposes to study a novel, underexplored, kind of visual relations that is also important to perceive and model -- inter-object functional relationships (e.g., a switch on the wall turns on or off the light, a remote control operates the TV). Humans often spend no effort or only a little to infer these relationships, even when entering a new room, by using our strong prior knowledge (e.g., we know that buttons control electrical devices) or using only a few exploratory interactions in cases of uncertainty (e.g., multiple switches and lights in the same room). In this paper, we take the first step in building AI system learning inter-object functional relationships in 3D indoor environments with key technical contributions of modeling prior knowledge by training over large-scale scenes and designing interactive policies for effectively exploring the training scenes and quickly adapting to novel test scenes. We create a new dataset based on the AI2Thor and PartNet datasets and perform extensive experiments that prove the effectiveness of our proposed method.

****

VAT-Mart: Learning Visual Action Trajectory Proposals for Manipulating 3D ARTiculated Objects

Ruihai Wu,Yan Zhao,Kaichun Mo,Zizheng Guo,Yian Wang,Tianhao Wu,Qingnan Fan,Xuelin Chen,Leonidas Guibas,Hao Dong

Perceiving and manipulating 3D articulated objects (e.g., cabinets, doors) in human environments is an important yet challenging task for future home-assistant robots. The space of 3D articulated objects is exceptionally rich in their myriad semantic categories, diverse shape geometry, and complicated part functionality. Previous works mostly abstract kinematic structure with estimated joint parameters and part poses as the visual representations for manipulating 3D articulated objects. In this paper, we propose object-centric actionable visual priors as a novel perception-interaction handshaking point that the perception system outputs more actionable guidance than kinematic structure estimation, by predicting dense geometry-aware, interaction-aware, and task-aware visual action affordance and trajectory proposals. We design an interaction-for-perception framework VAT-Mart to learn such actionable visual representations by simultaneously training a curiosity-driven reinforcement learning policy exploring diverse interaction trajectories and a perception module summarizing and generalizing the explored knowledge for pointwise predictions among diverse shapes. Experiments prove the effectiveness of the proposed approach using the large-scale PartNet-Mobility dataset in SAPIEN environment and show promising generalization capabilities to novel test shapes, unseen object categories, and real-world data.

****

Neural graphical modelling in continuous-time: consistency guarantees and algorithms

Alexis Bellot,Kim Branson,Mihaela van der Schaar

The discovery of structure from time series data is a key problem in fields of study working with complex systems. Most identifiability results and learning algorithms assume the underlying dynamics to be discrete in time. Comparatively few, in contrast, explicitly define dependencies in infinitesimal intervals of time, independently of the scale of observation and of the regularity of sampling. In this paper, we consider score-based structure learning for the study of dynamical systems. We prove that for vector fields parameterized in a large class of neural networks, least squares optimization with adaptive regularization schemes consistently recovers directed graphs of local independencies in systems of stochastic differential equations. Using this insight, we propose a score-based learning algorithm based on penalized Neural Ordinary Differential Equations (modelling the mean process) that we show to be applicable to the general setting of irregularly-sampled multivariate time series and to outperform the state of the art across a range of dynamical systems.

****

Kernel Deformed Exponential Families for Sparse Continuous Attention

Alexander Moreno,Supriya Nagesh,Zhenke Wu,Walter Dempsey,James Matthew Rehg

Attention mechanisms take an expectation of a data representation with respect to probability weights. This creates summary statistics that focus on important features. 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 use Gaussian mixture attention densities, which are a flexible class with dense support. In this paper, we extend this to two general flexible classes: kernel exponential families 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. Experiments show that kernel deformed exponential families can attend to non-overlapping intervals of time.

****

Hybrid Local SGD for Federated Learning with Heterogeneous Communications

Yuanxiong Guo,Ying Sun,Rui Hu,Yanmin Gong

Communication is a key bottleneck in federated learning where a large number of edge devices collaboratively learn a model under the orchestration of a central server without sharing their own training data. While local SGD has been proposed to reduce the number of FL rounds and become the algorithm of choice for FL, its total communication cost is still prohibitive when each device needs to communicate with the remote server repeatedly for many times over bandwidth-limited networks. In light of both device-to-device (D2D) and device-to-server (D2S) cooperation opportunities in modern communication networks, this paper proposes a new federated optimization algorithm dubbed hybrid local SGD (HL-SGD) in FL settings where devices are grouped into a set of disjoint clusters with high D2D communication bandwidth. HL-SGD subsumes previous proposed algorithms such as local SGD and gossip SGD and enables us to strike the best balance between model accuracy and runtime. We analyze the convergence of HL-SGD in the presence of heterogeneous data for general nonconvex settings. We also perform extensive experiments and show that the use of hybrid model aggregation via D2D and D2S communications in HL-SGD can largely speed up the training time of federated learning.

****

Recurrent Model-Free RL is a Strong Baseline for Many POMDPs

Tianwei Ni,Benjamin Eysenbach,Sergey Levine,Ruslan Salakhutdinov

Many problems in RL, such as meta RL, robust RL, and generalization in RL can be cast as POMDPs. In theory, simply augmenting model-free RL with memory, such as recurrent neural networks, provides a general approach to solving all types of POMDPs. However, prior work has found that such recurrent model-free RL methods tend to perform worse than more specialized algorithms that are designed for specific types of POMDPs. This paper revisits this claim. We find that a careful architecture and hyperparameter decisions yield a recurrent model-free implementation that performs on par with (and occasionally substantially better than) more sophisticated recent techniques in their respective domains. We also release a simple and efficient implementation of recurrent model-free RL for future work to use as a baseline for POMDPs.

****

C-Planning: An Automatic Curriculum for Learning Goal-Reaching Tasks

Tianjun Zhang,Benjamin Eysenbach,Ruslan Salakhutdinov,Sergey Levine,Joseph E. Gonzalez

Goal-conditioned reinforcement learning (RL) has shown great success recently at solving a wide range of tasks(e.g., navigation, robotic manipulation). However, learning to reach distant goals remains a central challenge to the field, and the task is particularly hard without any offline data, expert demonstrations, and reward shaping. In this paper, we propose to solve the distant goal-reaching task by using search at training time to generate a curriculum of intermediate states. Specifically, we introduce the algorithm Classifier-Planning (C-Planning) by framing the learning of the goal-conditioned policies as variational inference. C-Planning naturally follows expectation maximization (EM): the E-step corresponds to planning an optimal sequence of waypoints using graph search, while the M-step aims to learn a goal-conditioned policy to reach those waypoints. One essential difficulty of designing such an algorithm is accurately modeling the distribution over way-points to sample from. In C-Planning, we propose to sample the waypoints using contrastive methods to learn a value function. Unlike prior methods that combine goal-conditioned RL with graph search, ours performs search only during training and not testing, significantly decreasing the compute costs of deploying the learned policy. Empirically, we demonstrate that our method not only improves the sample efficiency of prior methods but also successfully solves temporally extended navigation and manipulation tasks, where prior goal-conditioned RL methods (including those based on graph search) fail to solve.

****

The Information Geometry of Unsupervised Reinforcement Learning

Benjamin Eysenbach,Ruslan Salakhutdinov,Sergey Levine

How can a reinforcement learning (RL) agent prepare to solve downstream tasks if those tasks are not known a priori? One approach is unsupervised skill discovery, a class of algorithms that learn a set of policies without access to a reward function. Such algorithms bear a close resemblance to representation learning a lgorithms (e.g., contrastive learning) in supervised learning, in that both are pretraining algorithms that maximize some approximation to a mutual information objective. While prior work has shown that the set of skills learned by such methods can accelerate downstream RL tasks, prior work offers little analysis into whether these skill learning algorithms are optimal, or even what notion of optimality would be appropriate to apply to them. In this work, we show that unsupervised skill discovery algorithms based on mutual information maximization do not learn skills that are optimal for every possible reward function. However, we show that the distribution over skills provides an optimal initialization minimizing regret against adversarially-chosen reward functions, assuming a certain type of adaptation procedure. Our analysis also provides a geometric perspective on these skill learning methods.

****

Mismatched No More: Joint Model-Policy Optimization for Model-Based RL
Benjamin Eysenbach,Alexander Khazatsky,Sergey Levine,Ruslan Salakhutdinov
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 increases a lower bound on expected return. This joint optimization mends the objective mismatch in prior work. Our objective is a global lower bound on expected return, and this bound becomes tight under certain assumptions. The resulting algorithm (MnM) is conceptually similar to 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.

****

Distinguishing rule- and exemplar-based generalization in learning systems
Ishita Dasgupta,Erin Grant,Thomas L. Griffiths
Despite the increasing scale of datasets in machine learning, generalization to unseen regions of the data distribution remains crucial. Such extrapolation is by definition underdetermined and is dictated by a learner’s inductive biases. Machine learning systems often do not share the same inductive biases as humans and, as a result, extrapolate in ways that are inconsistent with our expectations. We investigate two distinct such inductive biases: feature-level bias (differences in which features are more readily learned) and exemplar-vs-rule bias (differences in how these learned features are used for generalization). Exemplar- vs. rule-based generalization has been studied extensively in cognitive psychology, and in this work we present a protocol inspired by these experimental approaches for directly probing this trade-off in learning systems. The measures we propose characterize changes in extrapolation behavior when feature coverage is manipulated in a combinatorial setting. We present empirical results across a range of models and across both expository and real-world image and language domains. We demonstrate that measuring the exemplar-rule trade-off while controlling for feature-level bias provides a more complete picture of extrapolation behavior than existing formalisms. We find that most standard neural network models have a propensity towards exemplar-based extrapolation and discuss the implications of these findings for research on data augmentation, fairness, and systematic generalization.

****

NAS-Bench-Suite: NAS Evaluation is (Now) Surprisingly Easy
Yash Mehta,Colin White,Arber Zela,Arjun Krishnakumar,Guri Zabergja,Shakiba Moradian,Mahmoud Safari,Kaicheng Yu,Frank Hutter
The release of tabular benchmarks, such as NAS-Bench-101 and NAS-Bench-201, has

significantly lowered the computational overhead for conducting scientific research in neural architecture search (NAS). Although they have been widely adopted and used to tune real-world NAS algorithms, these benchmarks are limited to small search spaces and focus solely on image classification. Recently, several new NAS benchmarks have been introduced that cover significantly larger search spaces over a wide range of tasks, including object detection, speech recognition, and natural language processing. However, substantial differences among these NAS benchmarks have so far prevented their widespread adoption, limiting researchers to using just a few benchmarks. In this work, we present an in-depth analysis of popular NAS algorithms and performance prediction methods across 25 different combinations of search spaces and datasets, finding that many conclusions drawn from a few NAS benchmarks do \emph{not} generalize to other benchmarks. To help remedy this problem, we introduce \nasbs, a comprehensive and extensible collection of NAS benchmarks, accessible through a unified interface, created with the aim to facilitate reproducible, generalizable, and rapid NAS research. Our code is available at https://github.com/automl/naslib.

****

Machine Learning For Elliptic PDEs: Fast Rate Generalization Bound, Neural Scaling Law and Minimax Optimality

Yiping Lu,Haoxuan Chen,Jianfeng Lu,Lexing Ying,Jose Blanchet

In this paper, we study the statistical limits of deep learning techniques for solving elliptic partial differential equations (PDEs) from random samples using the Deep Ritz Method (DRM) and Physics-Informed Neural Networks (PINNs). To simplify the problem, we focus on a prototype elliptic PDE: the Schr\"odinger equation on a hypercube with zero Dirichlet boundary condition, which has wide application in the quantum-mechanical systems. We establish upper and lower bounds for both methods, which improves upon concurrently developed upper bounds for this problem via a fast rate generalization bound. We discover that the current Deep Ritz Methods is sub-optimal and propose a modified version of it. We also prove that PINN and the modified version of DRM can achieve minimax optimal bounds over Sobolev spaces. Empirically, following recent work which has shown that the deep model accuracy will improve with growing training sets according to a power law, we supply computational experiments to show a similar behavior of dimension dependent power law for deep PDE solvers.

****

MeshInversion: 3D textured mesh reconstruction with generative prior

Junzhe Zhang,Daxuan Ren,Zhongang Cai,Chai Kiat Yeo,Bo Dai,Chen Change Loy

Recovering a textured 3D mesh from a single image is highly challenging, particularly for in-the-wild objects that lack 3D ground truths. Prior attempts resort to weak supervision based on 2D silhouette annotations of monocular images. Since the supervision lies in the 2D space while the output is in the 3D space, such in-direct supervision often over-emphasizes the observable part of the 3D textured mesh, at the expense of the overall reconstruction quality. Although previous attempts have adopted various hand-crafted heuristics to reduce this gap, this issue is far from being solved. In this work, we present an alternative framework, \textbf{MeshInversion}, that reduces the gap by exploiting the \textit{generative prior} of a 3D GAN pre-trained for 3D textured mesh synthesis. Reconstruction is achieved by searching for a latent space in the 3D GAN that best resembles the target mesh in accordance with the single view observation. Since the pre-trained GAN encapsulates rich 3D semantics in terms of mesh geometry and texture, searching within the GAN manifold thus naturally regularizes the realness and fidelity of the reconstruction. Importantly, such regularization is directly applied in the 3D space, providing crucial guidance of mesh parts that are unobserved in the 2D space. Experiments on standard benchmarks show that our framework obtains faithful 3D reconstructions with consistent geometry and texture across both observed and unobserved parts. Moreover, it generalizes well to meshes that are less commonly seen, such as the extended articulation of deformable objects.

****

Characterising the Area Under the Curve Loss Function Landscape

Maximilian Paul Niroomand,Conor T Cafolla,John William Roger Morgan,David John W ales

One of the most common metrics to evaluate neural network classifiers is the area under the receiver operating characteristic curve (AUC). However, optimisation of the AUC as the loss function during network training is not a standard procedure. Here we compare minimising the cross-entropy (CE) loss and optimising the AUC directly. In particular, we analyse the loss function landscape (LFL) of approximate AUC (appAUC) loss functions to discover the organisation of this solution space. We discuss various surrogates for AUC approximation and show their differences. We find that the characteristics of the appAUC landscape are significantly different from the CE landscape. The approximate AUC loss function improves testing AUC, and the appAUC landscape has substantially more minima, but these minima are less robust, with larger average Hessian eigenvalues. We provide a theoretical foundation to explain these results. To generalise our results, we lastly provide an overview of how the LFL can help to guide loss function analysis and selection.

****

On the Capacity and Superposition of Minima in Neural Network Loss Function Landscapes

Maximilian Paul Niroomand,John William Roger Morgan,Conor T Cafolla,David John Wales

Minima of the loss function landscape of a neural network are locally optimal sets of weights that extract and process information from the input data to make outcome predictions. In underparameterised networks, the capacity of the weights may be insufficient to fit all the relevant information. We demonstrate that different local minima specialise in certain aspects of the learning problem, and process the input information differently. This effect can be exploited using a meta-network in which the predictive power from multiple minima of the LFL is combined to produce a better classifier. With this approach, we can increase the area under the receiver operating characteristic curve (AUC) by around 20% for a complex learning problem. We propose a theoretical basis for combining minima and show how a meta-network can be trained to select the representative that is used for classification of a specific data item. Finally, we present an analysis of symmetry-equivalent solutions to machine learning problems, which provides a systematic means to improve the efficiency of this approach.

****

Improving Hyperparameter Optimization by Planning Ahead

Hadi Samer Jomaa,Jonas Falkner,Lars Schmidt-Thieme

Hyperparameter optimization (HPO) is generally treated as a bi-level optimization problem that involves fitting a (probabilistic) surrogate model to a set of observed hyperparameter responses, e.g. validation loss, and consequently maximizing an acquisition function using a surrogate model to identify good hyperparameter candidates for evaluation. The choice of a surrogate and/or acquisition function can be further improved via knowledge transfer across related tasks. In this paper, we propose a novel transfer learning approach, defined within the context of model-based reinforcement learning, where we represent the surrogate as an ensemble of probabilistic models that allows trajectory sampling. We further propose a new variant of model predictive control which employs a simple look-ahead strategy as a policy that optimizes a sequence of actions, representing hyperparameter candidates to expedite HPO. Our experiments on three meta-datasets comparing to state-of-the-art HPO algorithms including a model-free reinforcement learning approach show that the proposed method can outperform all baselines by exp

loiting a simple planning-based policy.
****

Variational oracle guiding for reinforcement learning
Dongqi Han,Tadashi Kozuno,Xufang Luo,Zhao-Yun Chen,Kenji Doya,Yuqing Yang,Dongsheng Li

How to make intelligent decisions is a central problem in machine learning and artificial intelligence. Despite recent successes of deep reinforcement learning (RL) in various decision making problems, an important but under-explored aspect is how to leverage oracle observation (the information that is invisible during online decision making, but is available during offline training) to facilitate learning. For example, human experts will look at the replay after a Poker game, in which they can check the opponents' hands to improve their estimation of the opponents' hands from the visible information during playing. In this work, we study such problems based on Bayesian theory and derive an objective to leverage oracle observation in RL using variational methods. Our key contribution is to propose a general learning framework referred to as variational latent oracle guiding (VLOG) for DRL. VLOG is featured with preferable properties such as its robust and promising performance and its versatility to incorporate with any value-based DRL algorithm. We empirically demonstrate the effectiveness of VLOG in online and offline RL domains with tasks ranging from video games to a challenging tile-based game Mahjong. Furthermore, we publish the Mahjong environment and an offline RL dataset as a benchmark to facilitate future research on oracle guiding (https://github.com/Agony5757/mahjong).
****

Pessimistic Model Selection for Offline Deep Reinforcement Learning
Chao-Han Huck Yang,Zhengling Qi,Yifan Cui,Pin-Yu Chen

Deep Reinforcement Learning (DRL) has demonstrated great potentials in solving sequential decision making problems in many applications. Despite its promising performance, practical gaps exist when deploying DRL in real-world scenarios. One main barrier is the over-fitting issue that leads to poor generalizability of the policy learned by DRL. In particular, for offline DRL with observational data, model selection is a challenging task as there is no ground truth available for performance demonstration, in contrast with the online setting with simulated environments. In this work, we propose a pessimistic model selection (PMS) approach for offline DRL with a theoretical guarantee, which features a tuning-free framework for finding the best policy among a set of candidate models. Two refined approaches are also proposed to address the potential bias of DRL model in identifying the optimal policy. Numerical studies demonstrated the superior performance of our approach over existing methods.
****

Deep Fusion of Multi-attentive Local and Global Features with Higher Efficiency for Image Retrieval
Baorong Shi

Image retrieval is to search images similar to the given query image by extracting features. Previously, methods that firstly search by global features then re-rank images using local feature matching were proposed, which has an excellent performance on many datasets. However, their drawbacks are also obvious. For example, the local feature matching consumes time and space greatly, the re-ranking process weakens the influence of global features, and the local feature learning is not accurate enough and semantic enough because of the trivial design. In this work, we proposed a Unifying Global and Attention-based Local Features Retrieval method (referred to as UGALR), which is an end-to-end and single-stage pipeline. Particularly, UGALR benefits from two aspects: 1) it accelerates extraction speed and reduces memory consumption by removing the re-ranking process and learning local feature matching with convolutional neural networks instead of RANSAC algorithm; 2) it learns more accurate and semantic local information through combining spatial and channel attention with the aid of intermediate supervision. Experiments on Revisited Oxford and Paris datasets validate the effectiveness of our approach, and we achieved state-of-the-art performance compared to other popular methods. The codes will be available soon.

****

CDTrans: Cross-domain Transformer for Unsupervised Domain Adaptation

Tongkun Xu,Weihua Chen,Pichao WANG,Fan Wang,Hao Li,Rong Jin

Unsupervised domain adaptation (UDA) aims to transfer knowledge learned from a labeled source domain to a different unlabeled target domain. Most existing UDA methods focus on learning domain-invariant feature representation, either from the domain level or category level, using convolution neural networks (CNNs)-based frameworks. One fundamental problem for the category level based UDA is the production of pseudo labels for samples in target domain, which are usually too noisy for accurate domain alignment, inevitably compromising the UDA performance.

With the success of Transformer in various tasks, we find that the cross-attention in Transformer is robust to the noisy input pairs for better feature alignment, thus in this paper Transformer is adopted for the challenging UDA task. Specifically, to generate accurate input pairs, we design a two-way center-aware labeling algorithm to produce pseudo labels for target samples. Along with the pseudo labels, a weight-sharing triple-branch transformer framework is proposed to apply self-attention and cross-attention for source/target feature learning and source-target domain alignment, respectively.

Such design explicitly enforces the framework to learn discriminative domain-specific and domain-invariant representations simultaneously. The proposed method is dubbed CDTrans (cross-domain transformer), and it provides one of the first attempts to solve UDA tasks with a pure transformer solution. Experiments show that our proposed method achieves the best performance on public UDA datasets, e.g. VisDA-2017 and DomainNet. Code and models are available at https://github.com/CDTrans/CDTrans.

****

Beyond ImageNet Attack: Towards Crafting Adversarial Examples for Black-box Domains

Qilong Zhang,Xiaodan Li,YueFeng Chen,Jingkuan Song,Lianli Gao,Yuan He,Hui Xue'

Adversarial examples have posed a severe threat to deep neural networks due to their transferable nature. Currently, various works have paid great efforts to enhance the cross-model transferability, which mostly assume the substitute model is trained in the same domain as the target model.

However, in reality, the relevant information of the deployed model is unlikely to leak.

Hence, it is vital to build a more practical black-box threat model to overcome this limitation and evaluate the vulnerability of deployed models.

In this paper, with only the knowledge of the ImageNet domain, we propose a Beyond ImageNet Attack (BIA) to investigate the transferability towards black-box domains (unknown classification tasks). Specifically, we leverage a generative model to learn the adversarial function for disrupting low-level features of input images.

Based on this framework, we further propose two variants to narrow the gap between the source and target domains from the data and model perspectives, respectively. Extensive experiments on coarse-grained and fine-grained domains demonstrate the effectiveness of our proposed methods. Notably,

our methods outperform state-of-the-art approaches by up to 7.71\% (towards coarse-grained domains) and 25.91\% (towards fine-grained domains) on average. Our code is available at \url{https://github.com/Alibaba-AAIG/Beyond-ImageNet-Attack}.

****

Learning to Schedule Learning rate with Graph Neural Networks

Yuanhao Xiong,Li-Cheng Lan,Xiangning Chen,Ruochen Wang,Cho-Jui Hsieh

Recent decades have witnessed great development of stochastic optimization in training deep neural networks. Learning rate scheduling is one of the most important factors that influence the performance of stochastic optimizers like Adam. Traditional methods seek to find a relatively proper scheduling among a limited number of pre-defined rules and might not accommodate a particular target problem. Instead, we propose a novel Graph-Network-based Scheduler (GNS), aiming at learning a specific scheduling mechanism without restrictions to existing principles

. By constructing a directed graph for the underlying neural network of the target problem, GNS encodes current dynamics with a graph message passing network and trains an agent to control the learning rate accordingly via reinforcement learning. The proposed scheduler can capture the intermediate layer information while being able to generalize to problems of varying scales. Besides, an efficient reward collection procedure is leveraged to speed up training. We evaluate our framework on benchmarking datasets, Fashion-MNIST and CIFAR10 for image classification, and GLUE for language understanding. GNS shows consistent improvement over popular baselines when training CNN and Transformer models. Moreover, GNS demonstrates great generalization to different datasets and network structures.

****

MergeBERT: Program Merge Conflict Resolution via Neural Transformers

Alexey Svyatkovskiy,Todd Mytkowicz,Negar Ghorbani,Sarah Fakhoury,Elizabeth A Dinella,Christian Bird,Neel Sundaresan,Shuvendu Lahiri

Collaborative software development is an integral part of the modern software development life cycle, essential to the success of large-scale software projects. When multiple developers make concurrent changes around the same lines of code, a merge conflict may occur.

Such conflicts stall pull requests and continuous integration pipelines for hours to several days, seriously hurting developer productivity.

In this paper, we introduce MergeBERT, a novel neural program merge framework based on the token-level three-way differencing and a transformer encoder model. Exploiting restricted nature of merge conflict resolutions, we reformulate the task of generating the resolution sequence as a classification task over a set of primitive merge patterns extracted from real-world merge commit data.

Our model achieves 63--68\% accuracy of merge resolution synthesis, yielding nearly a 3$\times$ performance improvement over existing structured, and 2$\times$ improvement over neural program merge tools. Finally, we demonstrate that MergeBERT is sufficiently flexible to work with source code files in Java, JavaScript, TypeScript, and C\# programming languages, and can generalize zero-shot to unseen languages.

****

You May Need both Good-GAN and Bad-GAN for Anomaly Detection

Riqiang Gao,Zhoubing Xu,Guillaume Chabin,Awais Mansoor,Florin-Cristian Ghesu,Bogdan Georgescu,Bennett A. Landman,Sasa Grbic

Generative adversarial nets (GAN) have been successfully adapted for anomaly detection, where end-to-end anomaly scoring by so-called Bad-GAN has shown promising results. A Bad-GAN generates pseudo anomalies at the low-density area of inlier distribution, and thus the inlier/outlier distinction can be approximated. However, the generated pseudo anomalies from existing Bad-GAN approaches may (1) converge to certain patterns with limited diversity, and (2) differ from the real anomalies, making the anomaly detection hard to generalize. In this work, we propose a new model called Taichi-GAN to address the aforementioned issues of a conventional Bad-GAN. First, a new orthogonal loss is proposed to regularize the cosine distance of decentralized generated samples in a Bad-GAN. Second, we utilize few anomaly samples (when available) with a conventional GAN, i.e., so-called Good-GAN, to draw the generated pseudo anomalies closer to the real anomalies. Our Taichi-GAN incorporates Good-GAN and Bad-GAN in an adversarial manner; which generates pseudo anomalies that contributing to a more robust discriminator for anomaly scoring, and thus anomaly detection. Substantial improvements can be observed from our proposed model on multiple simulated and real-life anatomy detection tasks.

****

Prototype Based Classification from Hierarchy to Fairness

Mycal Tucker,Julie Shah

Artificial neural nets can represent and classify many types of high-dimensional data but are often tailored to particular applications -- e.g., for ``fair'' or ``hierarchical'' classification. Once an architecture has been selected, it is often difficult for humans to adjust models for a new task; for example, a hierarchical classifier cannot be easily transformed into a fair classifier that shields a protected field. Our contribution in this work is a new neural network architecture, the concept subspace network (CSN), which generalizes existing specialized classifiers to produce a unified model capable of learning a spectrum of multi-concept relationships. We demonstrate that CSNs reproduce state-of-the-art results in fair classification when enforcing concept independence, may be transformed into hierarchical classifiers, or may even reconcile fairness and hierarchy within a single classifier. The CSN is inspired by and matches the performance of existing prototype-based classifiers that promote interpretability.

****

SketchODE: Learning neural sketch representation in continuous time

Ayan Das,Yongxin Yang,Timothy Hospedales,Tao Xiang,Yi-Zhe Song

Learning meaningful representations for chirographic drawing data such as sketches, handwriting, and flowcharts is a gateway for understanding and emulating human creative expression. Despite being inherently continuous-time data, existing works have treated these as discrete-time sequences, disregarding their true nature. In this work, we model such data as continuous-time functions and learn compact representations by virtue of Neural Ordinary Differential Equations. To this end, we introduce the first continuous-time Seq2Seq model and demonstrate some remarkable properties that set it apart from traditional discrete-time analogues. We also provide solutions for some practical challenges for such models, including introducing a family of parameterized ODE dynamics & continuous-time data augmentation particularly suitable for the task. Our models are validated on several datasets including VectorMNIST, DiDi and Quick, Draw!.

****

Measuring the Interpretability of Unsupervised Representations via Quantized Reversed Probing

Iro Laina,Yuki M Asano,Andrea Vedaldi

Self-supervised visual representation learning has recently attracted significant research interest. While a common way to evaluate self-supervised representations is through transfer to various downstream tasks, we instead investigate the problem of measuring their interpretability, i.e. understanding the semantics encoded in raw representations. We formulate the latter as estimating the mutual information between the representation and a space of manually labelled concepts. To quantify this we introduce a decoding bottleneck: information must be captured by simple predictors, mapping concepts to clusters in representation space. This approach, which we call reverse linear probing, provides a single number sensitive to the semanticity of the representation. This measure is also able to detect when the representation contains combinations of concepts (e.g., "red apple'') instead of just individual attributes ("red'' and "apple'' independently). Finally, we propose to use supervised classifiers to automatically label large datasets in order to enrich the space of concepts used for probing. We use our method to evaluate a large number of self-supervised representations, ranking them by interpretability, highlight the differences that emerge compared to the standard evaluation with linear probes and discuss several qualitative insights. Code at: https://github.com/iro-cp/ssl-qrp.

****

Head2Toe: Utilizing Intermediate Representations for Better OOD Generalization

Utku Evci,Vincent Dumoulin,Hugo Larochelle,Michael Curtis Mozer

Transfer-learning methods aim to improve performance in a data-scarce target domain using a model pretrained on a data-rich source domain. A cost-efficient strategy, linear probing, involves freezing the source model and training a new classification head for the target domain. This strategy is outperformed by a more costly but state-of-the-art method---fine-tuning all parameters of the source model to the target domain---possibly because fine-tuning allows the model to leverage useful information from intermediate layers which is otherwise discarded by the later pretrained layers. We explore the hypothesis that these intermediate layers might be directly exploited by linear probing. We propose a method, Head-to-Toe probing (Head2Toe), that selects features from all layers of the source model to train a classification head for the target-domain. In evaluations on the VTAB, Head2Toe matches performance obtained with fine-tuning on average, but critically, for out-of-distribution transfer, Head2Toe outperforms fine-tuning.

****

GradMax: Growing Neural Networks using Gradient Information

Utku Evci,Bart van Merrienboer,Thomas Unterthiner,Fabian Pedregosa,Max Vladymyrov

The architecture and the parameters of neural networks are often optimized independently, which requires costly retraining of the parameters whenever the architecture is modified. In this work we instead focus on growing the architecture without requiring costly retraining. We present a method that adds new neurons during training without impacting what is already learned, while improving the training dynamics. We achieve the latter by maximizing the gradients of the new weights and efficiently find the optimal initialization by means of the singular value decomposition (SVD). We call this technique Gradient Maximizing Growth (GradMax) and demonstrate its effectiveness in variety of vision tasks and architectures. We open sourced our code at https://github.com/google-research/growneuron

****

Online Coreset Selection for Rehearsal-based Continual Learning

Jaehong Yoon,Divyam Madaan,Eunho Yang,Sung Ju Hwang

A dataset is a shred of crucial evidence to describe a task. However, each data point in the dataset does not have the same potential, as some of the data points can be more representative or informative than others. This unequal importance among the data points may have a large impact in rehearsal-based continual learning, where we store a subset of the training examples (coreset) to be replayed later to alleviate catastrophic forgetting. In continual learning, the quality of the samples stored in the coreset directly affects the model's effectiveness and efficiency. The coreset selection problem becomes even more important under realistic settings, such as imbalanced continual learning or noisy data scenarios. To tackle this problem, we propose Online Coreset Selection (OCS), a simple yet effective method that selects the most representative and informative coreset at each iteration and trains them in an online manner. Our proposed method maximizes the model's adaptation to a target dataset while selecting high-affinity samples to past tasks, which directly inhibits catastrophic forgetting. We validate the effectiveness of our coreset selection mechanism over various standard, imbalanced, and noisy datasets against strong continual learning baselines, demonstrating that it improves task adaptation and prevents catastrophic forgetting in a sample-efficient manner.

****

Switch to Generalize: Domain-Switch Learning for Cross-Domain Few-Shot Classification

Zhengdong Hu,Yifan Sun,Yi Yang

This paper considers few-shot learning under the cross-domain scenario. The cross-domain setting imposes a critical challenge, i.e., using very few (support) samples to generalize the already-learned model to a novel domain. We hold a hypothesis, i.e., if a deep model is capable to fast generalize itself to different domains (using very few samples) during training, it will maintain such domain generalization capacity for testing. It motivates us to propose a novel Domain-Switch Learning (DSL) framework. DSL embeds the cross-domain scenario into the training stage in a ``fast switching'' manner. Specifically, DSL uses a single domain for a training iteration and switches into another domain for the following iteration. During the switching, DSL enforces two constraints: 1) the deep model should not over-fit the domain in the current iteration and 2) the deep model should not forget the already-learned knowledge of other domains. These two constraints jointly promote fast generalization across different domains. Experimental results confirm that the cross-domain generalization capacity can be inherited from the training stage to the testing stage, validating our key hypothesis. Consequentially, DSL significantly improves cross-domain few-shot classification and sets up new state of the art.

****

Zero-CL: Instance and Feature decorrelation for negative-free symmetric contrastive learning

Shaofeng Zhang,Feng Zhu,Junchi Yan,Rui Zhao,Xiaokang Yang

For self-supervised contrastive learning, models can easily collapse and generate trivial constant solutions. The issue has been mitigated by recent improvement on objective design, which however often requires square complexity either for the size of instances ($\mathcal{O}(N^{2})$) or feature dimensions ($\mathcal{O}(d)^2$). To prevent such collapse, we develop two novel methods by decorrelating on different dimensions on the instance embedding stacking matrix, i.e., \textbf{I}nstance-wise (ICL) and \textbf{F}eature-wise (FCL) \textbf{C}ontrastive \textbf{L}earning. The proposed two methods (FCL, ICL) can be combined synthetically, called Zero-CL, where ``Zero'' means negative samples are \textbf{zero} relevant, which allows Zero-CL to completely discard negative pairs i.e., with \textbf{zero} negative samples. Compared with previous methods, Zero-CL mainly enjoys three advantages: 1) Negative free in symmetric architecture. 2) By whitening transformation, the correlation of the different features is equal to zero, alleviating information redundancy. 3) Zero-CL remains original information to a great extent after transformation, which improves the accuracy against other whitening transformation techniques. Extensive experimental results on CIFAR-10/100 and ImageNet show that Zero-CL outperforms or is on par with state-of-the-art symmetric contrastive learning methods.

****

Training-Free Robust Multimodal Learning via Sample-Wise Jacobian Regularization

Zhengqi Gao,Sucheng Ren,Zihui Xue,Siting Li,Hang Zhao

Multimodal fusion emerges as an appealing technique to improve model performances on many tasks. Nevertheless, the robustness of such fusion methods is rarely involved in the present literature. In this paper, we are the first to propose a training-free robust late-fusion method by exploiting conditional independence assumption and Jacobian regularization. Our key is to minimize the Frobenius norm of a Jacobian matrix, where the resulting optimization problem is relaxed to a tractable Sylvester equation. Furthermore, we provide a theoretical error bound of our method and some insights about the function of the extra modality. Several numerical experiments on AV-MNIST, RAVDESS, and VGGsound demonstrate the efficacy of our method under both adversarial attacks and random corruptions.

****

Stochastic Reweighted Gradient Descent

Ayoub El Hanchi,Chris J. Maddison,David Alan Stephens

Importance sampling is a promising strategy for improving the convergence rate of stochastic gradient methods. It is typically used to precondition the optimization problem, but it can also be used to reduce the variance of the gradient estimator. Unfortunately, this latter point of view has yet to lead to practical methods that improve the asymptotic error of stochastic gradient methods. In this work, we propose stochastic reweighted gradient (SRG), a variance-reduced stochastic gradient method based solely on importance sampling that can improve on the asymptotic error of stochastic gradient descent (SGD) in the strongly convex and smooth case. We show that SRG can be extended to combine the benefits of both importance-sampling-based preconditioning and variance reduction. When compared to SGD, the resulting algorithm can simultaneously reduce the condition number and the asymptotic error, both by up to a factor equal to the number of component functions. We demonstrate improved convergence in practice on ℓ_2-regularized logistic regression problems.

****

When less is more: Simplifying inputs aids neural network understanding

Robin Tibor Schirrmeister,Rosanne Liu,Sara Hooker,Tonio Ball

Are all bits useful? In this work, we propose SimpleBits, a method to synthesize simplified inputs by reducing information content, and carefully measure the effect of such simplification on learning. Crucially, SimpleBits does not require any domain-specific knowledge to constrain which input features should be removed. Instead, SimpleBits learns to remove the features of inputs which are least r elevant for a given task. Concretely, we jointly optimize for input simplification by reducing inputs' bits per dimension as given by a pretrained generative model, as well as for the classification performance. We apply the simplification approach to a wide range of scenarios: conventional training, dataset condensation and post-hoc explanations. In this way, we analyze what simplified inputs tell us about the decisions made by classification networks. We show that our simplification approach successfully removes superfluous information for tasks with injected distractors. When applied post-hoc, our approach provides intuition into reasons for misclassifications of conventionally trained classifiers. Finally, for dataset condensation, we find that inputs can be simplified with only minimal accuracy degradation. Overall, our learning-based simplification approach offers a valuable new tool to explore the basis of network decisions.

****

Random matrices in service of ML footprint: ternary random features with no performance loss

Hafiz Tiomoko Ali,Zhenyu Liao,Romain Couillet

In this article, we investigate the spectral behavior of random features kernel matrices of the type ${\bf K} = \mathbb{E}_{{\bf w}} \left[\sigma\left({\bf w}^{\sf T}{\bf x}_i\right)\sigma\left({\bf w}^{\sf T}{\bf x}_j\right)\right]_{i,j=1}^n$, with nonlinear function $\sigma(\cdot)$, data ${\bf x}_1, \ldots, {\bf x}_n \in \mathbb{R}^p$, and random projection vector ${\bf w} \in \mathbb{R}^p$ having i.i.d. entries. In a high-dimensional setting where the number of data n and their dimension p are both large and comparable, we show, under a Gaussian mixture model for the data, that the eigenspectrum of ${\bf K}$ is independent of the distribution of the i.i.d.(zero-mean and unit-variance) entries of ${\bf w}$, and only depends on $\sigma(\cdot)$ via its (generalized) Gaussian moments $\mathbb{E}_{z\sim \mathcal N(0,1)}[\sigma'(z)]$ and $\mathbb{E}_{z\sim \mathcal N(0,1)}[\sigma''(z)]$. As a result, for any kernel matrix ${\bf K}$ of the form above, we propose a novel random features technique, called Ternary Random Features (TRFs), that (i) asymptotically yields the same limiting kernel as the original ${\bf K}$ in a spectral sense and (ii) can be computed and stored much more efficiently, by wisely tuning (in a data-dependent manner) the function σ and the random vector ${\bf w}$, both taking values in $\{-1,0,1\}$. The computation of the proposed random features requires no multiplication, and a factor of b times less bits for storage compared to classical random features such as random Fourier features, with b the number of bits to store full precision values. Besides, it appears in our experiments on real data that the substantial gains in computation and storage are accompanied with somewhat improved performances compared to state-of-the-art random features methods.

****

Kalman Filter Is All You Need: Optimization Works When Noise Estimation Fails

Ido Greenberg,Shie Mannor,Netanel Yannay

Determining the noise parameters of a Kalman Filter (KF) has been studied for decades. A huge body of research focuses on the task of noise estimation under various conditions, since precise noise estimation is considered equivalent to minimization of the filtering errors. However, we show that even a small violation of the KF assumptions can significantly modify the effective noise, breaking the equivalence between the tasks and making noise estimation an inferior strategy. We show that such violations are common, and are often not trivial to handle or even notice. Consequentially, we argue that a robust solution is needed - rather than choosing a dedicated model per problem.

To that end, we apply gradient-based optimization to the filtering errors directly, with relation to an efficient parameterization of the symmetric and positive-definite parameters of the KF. In a variety of state-estimation and tracking problems, we show that the optimization improves both the accuracy of the KF and its robustness to design decisions.

In addition, we demonstrate how an optimized neural network model can seem to reduce the errors significantly compared to a KF - and how this reduction vanishes once the KF is optimized similarly. This indicates how complicated models can be wrongly identified as superior to the KF, while in fact they were merely more optimized.

****

Global Convergence of Multi-Agent Policy Gradient in Markov Potential Games

Stefanos Leonardos,Will Overman,Ioannis Panageas,Georgios Piliouras

Potential games are arguably one of the most important and widely studied classes of normal form games. They define the archetypal setting of multi-agent coordination in which all agents utilities are perfectly aligned via a common potential function. Can this intuitive framework be transplanted in the setting of Markov games? What are the similarities and differences between multi-agent coordination with and without state dependence? To answer these questions, we study a natural class of Markov Potential Games (MPGs) that generalize prior attempts at capturing complex stateful multi-agent coordination. Counter-intuitively, insights from normal-form potential games do not carry over as MPGs involve settings where state-games can be zero-sum games. In the opposite direction, Markov games where every state-game is a potential game are not necessarily MPGs. Nevertheless, MPGs showcase standard desirable properties such as the existence of deterministic Nash policies. In our main technical result, we prove convergence of independent policy gradient and its stochastic counterpart to Nash policies (polynomially fast in the approximation error) by adapting recent gradient dominance property arguments developed for single-agent Markov decision processes to multi-agent learning settings.

****

Rethinking Adversarial Transferability from a Data Distribution Perspective

Yao Zhu,Jiacheng Sun,Zhenguo Li

Adversarial transferability enables attackers to generate adversarial examples from the source model to attack the target model, which has raised security concerns about the deployment of DNNs in practice. In this paper, we rethink adversarial transferability from a data distribution perspective and further enhance transferability by score matching based optimization. We identify that some samples with injecting small Gaussian noise can fool different target models, and their adversarial examples under different source models have much stronger transferability. We hypothesize that these samples are in the low-density region of the ground truth distribution where models are not well trained. To improve the attack success rate of adversarial examples, we match the adversarial attacks with the directions which effectively decrease the ground truth density. We propose Intrinsic Adversarial Attack (IAA), which smooths the activation function and decreases the impact of the later layers of a given normal model, to increase the alignment of adversarial attack and the gradient of joint data distribution. We conduct comprehensive transferable attacks against multiple DNNs and show that our IAA can boost the transferability of the crafted attacks in all cases and go beyond state-of-the-art methods.

****

Dynamic Parameterized Network for CTR Prediction

Jian Zhu,Congcong Liu,Pei Wang,Xiwei Zhao,Guangpeng Chen,Jin Jun Sheng,Changping Peng,Zhangang Lin,Jingping Shao

Learning to capture feature relations effectively and efficiently is essential in click-through rate (CTR) prediction of modern recommendation systems. Most existing CTR prediction methods model such relations either through tedious manually-designed low-order interactions or through inflexible and inefficient high-order interactions, which both require extra DNN modules for implicit interaction modeling. In this paper, we proposed a novel plug-in operation, Dynamic Parameterized Operation (DPO), to learn both explicit and implicit interaction instance-wisely. We showed that the introduction of DPO into DNN modules and Attention modules can respectively benefit two main tasks in CTR prediction, enhancing the adaptiveness of feature-based modeling and improving user behavior modeling with the instance-wise locality. Our Dynamic Parameterized Networks significantly outperforms state-of-the-art methods in the offline experiments on the public dataset and real-world production dataset, together with an online A/B test. Furthermore, the proposed Dynamic Parameterized Networks has been deployed in the ranking system of one of the world's largest e-commerce companies, serving the main traffic of hundreds of millions of active users.

****

Towards Understanding the Condensation of Neural Networks at Initial Training

Zhiqin Xu,Hanxu Zhou,Tao Luo,Yaoyu Zhang

Implicit regularization is important for understanding the learning of neural networks (NNs). Empirical works show that 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 on isolated orientations with a small initialization. The condensation dynamics implies that the training implicitly regularizes a NN towards one with much smaller effective size. In this work, we utilize multilayer networks to show that the maximal number of condensed orientations in the initial training stage is twice the multiplicity of the activation function, where ``multiplicity'' is multiple roots of activation function at origin. Our theoretical analysis confirms experiments for two cases, one is for the activation function of multiplicity one, which contains many common activation functions, and the other is for the layer with one-dimensional input. This work makes a step towards understanding how small initialization implicitly leads NNs to condensation at initial training stage, which lays a foundation for the future study of the nonlinear dynamics of NNs and its implicit regularization effect at a later stage of training.

****

Representation Disentanglement in Generative Models with Contrastive Learning

Shentong Mo,Zhun Sun,Shumin Han

Contrastive learning has shown its effectiveness in image classification and generation. Recent works apply the contrastive learning on the discriminator of the Generative Adversarial Networks, and there exists little work on exploring if contrastive learning can be applied on encoders to learn disentangled representations. In this work, we propose a simple yet effective method via incorporating contrastive learning into latent optimization, where we name it $\textbf{\texttt{ContraLORD}}$. Specifically, we first use a generator to learn discriminative and disentangled embeddings via latent optimization. Then an encoder and two momentum encoders are applied to dynamically learn disentangled information across large amount of samples with content-level and residual-level contrastive loss. In the meanwhile, we tune the encoder with the learned embeddings in an amortized manner. We evaluate our approach on ten benchmarks in terms of representation disentanglement and linear classification. Extensive experiments demonstrate the effectiveness of our ContraLORD on learning both discriminative and generative representations.

****

Transformers Can Do Bayesian Inference

Samuel Müller,Noah Hollmann,Sebastian Pineda Arango,Josif Grabocka,Frank Hutter

Currently, it is hard to reap the benefits of deep learning for Bayesian methods, which allow the explicit specification of prior knowledge and accurately capture model uncertainty. We present Prior-Data Fitted Networks (PFNs). PFNs leverage large-scale machine learning techniques to approximate a large set of posteriors. The only requirement for PFNs to work is the ability to sample from a prior distribution over supervised learning tasks (or functions). Our method restates the objective of posterior approximation as a supervised classification problem with a set-valued input: it repeatedly draws a task (or function) from the prior, draws a set of data points and their labels from it, masks one of the labels and learns to make probabilistic predictions for it based on the set-valued input of the rest of the data points. Presented with a set of samples from a new supervised learning task as input, PFNs make probabilistic predictions for arbitrary other data points in a single forward propagation, having learned to approximate Bayesian inference. We demonstrate that PFNs can near-perfectly mimic Gaussian processes and also enable efficient Bayesian inference for intractable problems, with over 200-fold speedups in multiple setups compared to current methods. We obtain strong results in very diverse areas such as Gaussian process regression, Bayesian neural networks, classification for small tabular data sets, and few-shot image classification, demonstrating the generality of PFNs. Code and trained PFNs are released at https://github.com/automl/TransformersCanDoBayesianInference.

****

Successive POI Recommendation via Brain-inspired Spatiotemporal Aware Representation

Gehua Ma,Jingyuan Zhao,Huajin Tang

POI vector representation (embedding) is the core of successive POI recommendation. However, existing approaches only rely on basic discretization and interval analyses and fail to fully exploit complicated spatiotemporal attributes of POIs. Neuroscience research has shown that the mammalian brain entorhinal-hippocampal system provides efficient graph representations for general knowledge. Moreover, entorhinal grid cells present concise spatial representations, while hippocampal place cells represent perception conjunctions effectively. Thus, the entorhinal-hippocampal system provides a novel angle for spatiotemporal aware representation, which inspires us to propose the SpatioTemporal aware Embedding framework (STE) and apply to POIs (STEP). STEP considers two types of POI-specific representations: sequential representation and spatiotemporal conjunctive representation, learned using sparse unlabeled data based on the proposed graph-building policies. Notably, the spatiotemporal conjunctive representation represents POIs from spatial and temporal aspects jointly and precisely. Furthermore, we introduce a user privacy secure successive POI recommendation method using STEP. Experimental results on two datasets demonstrate that STEP captures POI-specific spatiotemporal information more accurately and achieves the state-of-the-art successive POI recommendation performance. Therefore, this work provides a novel solution to spatiotemporal aware representation and paves a new way for spatiotemporal modeling-related tasks.

****

Learning Discrete Structured Variational Auto-Encoder using Natural Evolution Strategies

Alon Berliner,Guy Rotman,Yossi Adi,Roi Reichart,Tamir Hazan

Discrete variational auto-encoders (VAEs) are able to represent semantic latent spaces in generative learning. In many real-life settings, the discrete latent space consists of high-dimensional structures, and propagating gradients through the relevant structures often requires enumerating over an exponentially large latent space. Recently, various approaches were devised to propagate approximated gradients without enumerating over the space of possible structures. In this work, we use Natural Evolution Strategies (NES), a class of gradient-free black-box optimization algorithms, to learn discrete structured VAEs. The NES algorithms are computationally appealing as they estimate gradients with forward pass evaluations only, thus they do not require to propagate gradients through their discrete structures. We demonstrate empirically that optimizing discrete structured VAEs using NES is as effective as gradient-based approximations. Lastly, we prove NES converges for non-Lipschitz functions as appear in discrete structured VAEs.

****

Learning Features with Parameter-Free Layers

Dongyoon Han,YoungJoon Yoo,Beomyoung Kim,Byeongho Heo

Trainable layers such as convolutional building blocks are the standard network design choices by learning parameters to capture the global context through successive spatial operations. When designing an efficient network, trainable layers such as the depthwise convolution is the source of efficiency in the number of parameters and FLOPs, but there was little improvement to the model speed in practice. This paper argues that simple built-in parameter-free operations can be a favorable alternative to the efficient trainable layers replacing spatial operations in a network architecture. We aim to break the stereotype of organizing the spatial operations of building blocks into trainable layers. Extensive experimental analyses based on layer-level studies with fully-trained models and neural architecture searches are provided to investigate whether parameter-free operations such as the max-pool are functional. The studies eventually give us a simpl e yet effective idea for redesigning network architectures, where the parameter-free operations are heavily used as the main building block without sacrificing the model accuracy as much. Experimental results on the ImageNet dataset demonstrate that the network architectures with parameter-free operations could enjoy the advantages of further efficiency in terms of model speed, the number of the parameters, and FLOPs. Code and ImageNet pretrained models are available at https://github.com/naver-ai/PfLayer.

****

Denoising Likelihood Score Matching for Conditional Score-based Data Generation

Chen-Hao Chao,Wei-Fang Sun,Bo-Wun Cheng,Yi-Chen Lo,Chia-Che Chang,Yu-Lun Liu,Yu-Lin Chang,Chia-Ping Chen,Chun-Yi Lee

Many existing conditional score-based data generation methods utilize Bayes' theorem to decompose the gradients of a log posterior density into a mixture of scores. These methods facilitate the training procedure of conditional score models, as a mixture of scores can be separately estimated using a score model and a classifier. However, our analysis indicates that the training objectives for the classifier in these methods may lead to a serious score mismatch issue, which corresponds to the situation that the estimated scores deviate from the true ones. Such an issue causes the samples to be misled by the deviated scores during the diffusion process, resulting in a degraded sampling quality. To resolve it, we theoretically formulate a novel training objective, called Denoising Likelihood Score Matching (DLSM) loss, for the classifier to match the gradients of the true log likelihood density. Our experimental evidences show that the proposed method outperforms the previous methods on both Cifar-10 and Cifar-100 benchmarks noticeably in terms of several key evaluation metrics. We thus conclude that, by adopting DLSM, the conditional scores can be accurately modeled, and the effect of the score mismatch issue is alleviated.

****

Language modeling via stochastic processes

Rose E Wang,Esin Durmus,Noah Goodman,Tatsunori Hashimoto

Modern language models can generate high-quality short texts. However, they often meander or are incoherent when generating longer texts. These issues arise from the next-token-only language modeling objective. To address these issues, we introduce Time Control (TC), a language model that implicitly plans via a latent stochastic process. TC does this by learning a representation which maps the dynamics of how text changes in a document to the dynamics of a stochastic process of interest. Using this representation, the language model can generate text by first implicitly generating a document plan via a stochastic process, and then generating text that is consistent with this latent plan. Compared to domain-specific methods and fine-tuning GPT2 across a variety of text domains, TC improves performance on text infilling and discourse coherence. On long text generation settings, TC preserves the text structure both in terms of ordering (up to +40% better) and text length consistency (up to +17% better). Human evaluators also prefer TC's output 28.6% more than the baselines.

****

A Study of Face Obfuscation in ImageNet

Kaiyu Yang,Jacqueline Yau,Li Fei-Fei,Jia Deng,Olga Russakovsky

Face obfuscation (blurring, mosaicing, etc.) has been shown to be effective for privacy protection; nevertheless, object recognition research typically assumes access to complete, unobfuscated images. In this paper, we explore the effects of face obfuscation on the popular ImageNet challenge visual recognition benchmark. Most categories in the ImageNet challenge are not people categories; however, many incidental people appear in the images, and their privacy is a concern. We first annotate faces in the dataset. Then we demonstrate that face blurring and overlaying---two typical obfuscation techniques---have minimal impact on the accuracy of recognition models. Concretely, we benchmark multiple deep neural networks on face-obfuscated images and observe that the overall recognition accuracy drops only slightly (<= 1.0%). Further, we experiment with transfer learning to

4 downstream tasks (object recognition, scene recognition, face attribute classification, and object detection) and show that features learned on face-obfuscated images are equally transferable. Our work demonstrates the feasibility of privacy-aware visual recognition, improves the highly-used ImageNet challenge benchmark, and suggests an important path for future visual datasets.

****

Learning Symbolic Rules for Reasoning in Quasi-Natural Language
Kaiyu Yang,Jia Deng

Symbolic reasoning, rule-based symbol manipulation, is a hallmark of human intelligence. However, rule-based systems have had limited success competing with learning-based systems outside formalized domains such as automated theorem proving. We hypothesize that this is due to the manual construction of rules in past attempts. In this work, we ask how we can build a rule-based system that can reason with natural language input but without the manual construction of rules. We propose MetaQNL, a "Quasi-Natural" language that can express both formal logic and natural language sentences, and MetaInduce, a learning algorithm that induces MetaQNL rules from training data consisting of questions and answers, with or without intermediate reasoning steps. Our approach achieves state-of-the-art accuracy on multiple reasoning benchmarks; it learns compact models with much less data and produces not only answers but also checkable proofs.

****

Memory Replay with Data Compression for Continual Learning
Liyuan Wang,Xingxing Zhang,Kuo Yang,Longhui Yu,Chongxuan Li,Lanqing HONG,Shifeng Zhang,Zhenguo Li,Yi Zhong,Jun Zhu

Continual learning needs to overcome catastrophic forgetting of the past. Memory replay of representative old training samples has been shown as an effective solution, and achieves the state-of-the-art (SOTA) performance. However, existing work is mainly built on a small memory buffer containing a few original data, which cannot fully characterize the old data distribution. In this work, we propose memory replay with data compression to reduce the storage cost of old training samples and thus increase their amount that can be stored in the memory buffer. Observing that the trade-off between the quality and quantity of compressed data is highly nontrivial for the efficacy of memory replay, we propose a novel method based on determinantal point processes (DPPs) to efficiently determine an appropriate compression quality for currently-arrived training samples. In this way, using a naive data compression algorithm with a properly selected quality can largely boost recent strong baselines by saving more compressed data in a limited storage space. We extensively validate this across several benchmarks of class-incremental learning and in a realistic scenario of object detection for autonomous driving.

****

MAML is a Noisy Contrastive Learner in Classification
Chia Hsiang Kao,Wei-Chen Chiu,Pin-Yu Chen

Model-agnostic meta-learning (MAML) is one of the most popular and widely adopted meta-learning algorithms, achieving remarkable success in various learning problems. Yet, with the unique design of nested inner-loop and outer-loop updates, which govern the task-specific and meta-model-centric learning, respectively, the underlying learning objective of MAML remains implicit, impeding a more straightforward understanding of it. In this paper, we provide a new perspective of the working mechanism of MAML. We discover that MAML is analogous to a meta-learner using a supervised contrastive objective in classification. The query features are pulled towards the support features of the same class and against those of different classes. Such contrastiveness is experimentally verified via an analysis based on the cosine similarity. Moreover, we reveal that vanilla MAML has an undesirable interference term originating from the random initialization and the cross-task interaction. We thus propose a simple but effective technique, the zeroing trick, to alleviate the interference. Extensive experiments are conducted on both mini-ImageNet and Omniglot datasets to validate the consistent improvement brought by our proposed method.

****

Noise Reconstruction and Removal Network: A New Way to Denoise FIB-SEM Images

Katya Giannios,Abhishek Chaurasia,Bambi DeLaRosa,Guillaume THIBAULT,Jessica L. Riesterer,Erin S Stempinski,Terence P Lo,Joe W Gray

Recent advances in Focused Ion Beam-Scanning Electron Microscopy (FIB-SEM) allow the imaging and analysis of cellular ultrastructure at nanoscale resolution, but the collection of labels and/or noise-free data sets has several challenges, often immutable. Reasons range from time consuming manual annotations, requiring highly trained specialists, to introducing imaging artifacts from the prolonged scanning during acquisition. We propose a fully unsupervised Noise Reconstruction and Removal Network for denoising scanning electron microscopy images. The architecture, inspired by gated recurrent units, reconstructs and removes the noise by synthesizing the sequential data. At the same time, the fully unsupervised training guides the network in distinguishing true signal from noise and gives comparable/even better results than supervised approaches on 3D electron microscopy data sets. We provide detailed performance analysis using numerical as well as empirical metrics.

****

RelViT: Concept-guided Vision Transformer for Visual Relational Reasoning

Xiaojian Ma,Weili Nie,Zhiding Yu,Huaizu Jiang,Chaowei Xiao,Yuke Zhu,Song-Chun Zhu,Anima Anandkumar

Reasoning about visual relationships is central to how humans interpret the visual world. This task remains challenging for current deep learning algorithms since it requires addressing three key technical problems jointly: 1) identifying object entities and their properties, 2) inferring semantic relations between pairs of entities, and 3) generalizing to novel object-relation combinations, i.e., systematic generalization. In this work, we use vision transformers (ViTs) as our base model for visual reasoning and make better use of concepts defined as object entities and their relations to improve the reasoning ability of ViTs. Specifically, we introduce a novel concept-feature dictionary to allow flexible image feature retrieval at training time with concept keys. This dictionary enables two new concept-guided auxiliary tasks: 1) a global task for promoting relational reasoning, and 2) a local task for facilitating semantic object-centric correspondence learning. To examine the systematic generalization of visual reasoning models, we introduce systematic splits for the standard HICO and GQA benchmarks. We show the resulting model, Concept-guided Vision Transformer (or RelViT for short) significantly outperforms prior approaches on HICO and GQA by 16% and 13% in the original split, and by 43% and 18% in the systematic split. Our ablation analyses also reveal our model's compatibility with multiple ViT variants and robustness to hyper-parameters.

****

Local Augmentation for Graph Neural Networks

Songtao Liu,Hanze Dong,Lanqing Li,Tingyang Xu,Yu Rong,Peilin Zhao,Junzhou Huang,Dinghao Wu

Data augmentation has been widely used in image data and linguistic data but remains under-explored on graph-structured data. Existing methods focus on augmenting the graph data from a global perspective and largely fall into two genres: structural manipulation and adversarial training with feature noise injection. However, the structural manipulation approach suffers information loss issues while the adversarial training approach may downgrade the feature quality by injecting noise. In this work, we introduce the local augmentation, which enhances node features by its local subgraph structures. Specifically, we model the data argumentation as a feature generation process. Given the central node's feature, our local augmentation approach learns the conditional distribution of its neighbors' features and generates the neighbors' optimal feature to boost the performance of downstream tasks. Based on the local augmentation, we further design a novel framework: LA-GNN, which can apply to any GNN models in a plug-and-play manner. Extensive experiments and analyses show that local augmentation consistently yields performance improvement for various GNN architectures across a diverse set of benchmarks.

****

Private Multi-Winner Voting For Machine Learning
Adam Dziedzic,Christopher A. Choquette-Choo,Natalie Dullerud,Vinith Menon Suriyakumar,Ali Shahin Shamsabadi,Muhammad Ahmad Kaleem,Somesh Jha,Nicolas Papernot,Xiao Wang

Private multi-winner voting is the task of revealing k-hot binary vectors that satisfy a bounded differential privacy guarantee. This task has been understudied in the machine learning literature despite its prevalence in many domains such as healthcare. We propose three new privacy-preserving multi-label mechanisms: Binary, τ, and Powerset voting. Binary voting operates independently per label through composition. τ voting bounds votes optimally in their ℓ_2 norm. Powerset voting operates over the entire binary vector by viewing the possible outcomes as a power set. We theoretically analyze tradeoffs showing that Powerset voting requires strong correlations between labels to outperform Binary voting. We use these mechanisms to enable privacy-preserving multi-label learning by extending the canonical single-label technique: PATE. We empirically compare our techniques with DPSGD on large real-world healthcare data and standard multi-label benchmarks. We find that our techniques outperform all others in the centralized setting. We enable multi-label CaPC and show that our mechanisms can be used to collaboratively improve models in a multi-site (distributed) setting.

****

Increasing the Cost of Model Extraction with Calibrated Proof of Work
Adam Dziedzic,Muhammad Ahmad Kaleem,Yu Shen Lu,Nicolas Papernot

In model extraction attacks, adversaries can steal a machine learning model exposed via a public API by repeatedly querying it and adjusting their own model based on obtained predictions. To prevent model stealing, existing defenses focus on detecting malicious queries, truncating, or distorting outputs, thus necessarily introducing a tradeoff between robustness and model utility for legitimate users. Instead, we propose to impede model extraction by requiring users to complete a proof-of-work before they can read the model's predictions. This deters attackers by greatly increasing (even up to 100x) the computational effort needed to leverage query access for model extraction. Since we calibrate the effort required to complete the proof-of-work to each query, this only introduces a slight overhead for regular users (up to 2x). To achieve this, our calibration applies tools from differential privacy to measure the information revealed by a query. Our method requires no modification of the victim model and can be applied by machine learning practitioners to guard their publicly exposed models against being easily stolen.

****

A NEW BACKBONE FOR HYPERSPECTRAL IMAGE RECONSTRUCTION
Jiamian Wang,Yulun Zhang,Xin Yuan,Yun Fu,ZHIQIANG TAO

As the inverse process of snapshot compressive imaging, the hyperspectral image (HSI) reconstruction takes the 2D measurement as input and posteriorly retrieves the captured 3D spatial-spectral signal. Built upon several assumptions, numerous sophisticated neural networks have come to the fore in this task. Despite their prosperity under experimental settings, it's still extremely challenging for existing networks to achieve high-fidelity reconstructive quality while maximizing the reconstructive efficiency (computational efficiency and power occupation), which prohibits their further deployment in practical applications. In this paper, we firstly conduct a retrospective analysis on aforementioned assumptions, through which we indicate the imminent aspiration for an authentically practical-oriented network in reconstructive community. By analysing the effectiveness and limitations of the widely-used reconstructive backbone U-Net, we propose a Simple Reconstruction Network, namely SRN, just based on some popular techniques, e.g., scale/spectral-invariant learning and identity connection. It turns out, under current conditions, such a pragmatic solution outperforms existing reconstructive methods by an obvious margin and maximize the reconstructive efficiency concretely. We hope the proposed SRN can further contribute to the cutting-edge reconstructive methods as a promising backbone, and also benefit the realistic tasks, i.e., real-time/high-resolution HSI reconstruction, solely as a baseline.

****

Robust Cross-Modal Semi-supervised Few Shot Learning

Xu Chen

Semi-supervised learning has been successfully applied to few-shot learning (FSL) due to its capability of leveraging the information of limited labeled data and massive unlabeled data. However, in many realistic applications, the query and support sets provided for FSL are potentially noisy or unreadable where the noise exists in both corrupted labels and outliers. Motivated by that, we propose to employ a robust cross-modal semi-supervised few-shot learning (RCFSL) based on Bayesian deep learning. By placing the uncertainty prior on top of the parameters of infinite Gaussian mixture model for noisy input, multi-modality information from image and text data are integrated into a robust heterogenous variational autoencoder. Subsequently, a robust divergence measure is employed to further enhance the robustness, where a novel variational lower bound is derived and optimized to infer the network parameters. Finally, a robust semi-supervised generative adversarial network is employed to generate robust features to compensate data sparsity in few shot learning and a joint optimization is applied for training and inference. Our approach is more parameter-efficient, scalable and adaptable compared to previous approaches. Superior performances over the state-of-the-art on multiple benchmark multi-modal dataset are demonstrated given the complicated noise for semi-supervised few-shot learning.

****

Adaptive Activation-based Structured Pruning

Kaiqi Zhao,Animesh Jain,Ming Zhao

Pruning is a promising approach to compress complex deep learning models in order to deploy them on resource-constrained edge devices. However, many existing pruning solutions are based on unstructured pruning, which yield models that cannot efficiently run on commodity hardware, and require users to manually explore and tune the pruning process, which is time consuming and often leads to sub-optimal results. To address these limitations, this paper presents an adaptive, activation-based, structured pruning approach to automatically and efficiently generate small, accurate, and hardware-efficient models that meet user requirements. First, it proposes iterative structured pruning using activation-based attention feature maps to effectively identify and prune unimportant filters. Then, it proposes adaptive pruning policies for automatically meeting the pruning objectives of accuracy-critical, memory-constrained, and latency-sensitive tasks. A comprehensive evaluation shows that the proposed method can substantially outperform the state-of-the-art structured pruning works on CIFAR-10 and ImageNet datasets. For example, on ResNet-56 with CIFAR-10, without any accuracy drop, our method achieves the largest parameter reduction (79.11%), outperforming the related works by 22.81% to 66.07%, and the largest FLOPs reduction (70.13%), outperforming the related works by 14.13% to 26.53%.

****

Boosted Curriculum Reinforcement Learning

Pascal Klink,Carlo D'Eramo,Jan Peters,Joni Pajarinen

Curriculum value-based reinforcement learning (RL) solves a complex target task by reusing action-values across a tailored sequence of related tasks of increasing difficulty. However, finding an exact way of reusing action-values in this setting is still a poorly understood problem. In this paper, we introduce the concept of boosting to curriculum value-based RL, by approximating the action-value function as a sum of residuals trained on each task. This approach, which we refer to as boosted curriculum reinforcement learning (BCRL), has the benefit of naturally increasing the representativeness of the functional space by adding a new residual each time a new task is presented. This procedure allows reusing previous action-values while promoting expressiveness of the action-value function.

We theoretically study BCRL as an approximate value iteration algorithm, discussing advantages over regular curriculum RL in terms of approximation accuracy and convergence to the optimal action-value function. Finally, we provide detailed empirical evidence of the benefits of BCRL in problems requiring curricula for accurate action-value estimation and targeted exploration.

****

Stochastic Deep Networks with Linear Competing Units for Model-Agnostic Meta-Learning

Konstantinos I. Kalais,Sotirios Chatzis

This work addresses meta-learning (ML) by considering deep networks with stochastic local winner-takes-all (LWTA) activations. This type of network units result in sparse representations from each model layer, as the units are organized into blocks where only one unit generates a non-zero output. The main operating principle of the introduced units lies on stochastic arguments, as the network performs posterior sampling over competing units to select the winner. Therefore, the proposed networks are explicitly designed to extract input data representations of sparse stochastic nature, as opposed to the currently standard deterministic representation paradigm. We posit that these modeling arguments, inspired from Bayesian statistics, allow for more robust modeling when uncertainty is high due to the limited availability of task-related training data; this is exactly the case with ML, which is the focus of this work. At training time, we rely on the reparameterization trick for Discrete distributions to perform reliable training via Monte-Carlo sampling. At inference time, we rely on Bayesian Model Averaging, which effectively averages over a number of sampled representations. As we experimentally show, our approach produces state-of-the-art predictive accuracy on standard few-shot image classification benchmarks; this is achieved without compromising computational efficiency.

****

ViDT: An Efficient and Effective Fully Transformer-based Object Detector

Hwanjun Song,Deqing Sun,Sanghyuk Chun,Varun Jampani,Dongyoon Han,Byeongho Heo,Wonjae Kim,Ming-Hsuan Yang

Transformers are transforming the landscape of computer vision, especially for recognition tasks. Detection transformers are the first fully end-to-end learning systems for object detection, while vision transformers are the first fully transformer-based architecture for image classification. In this paper, we integrate Vision and Detection Transformers (ViDT) to build an effective and efficient object detector. ViDT introduces a reconfigured attention module to extend the recent Swin Transformer to be a standalone object detector, followed by a computationally efficient transformer decoder that exploits multi-scale features and auxiliary techniques essential to boost the detection performance without much increase in computational load. Extensive evaluation results on the Microsoft COCO benchmark dataset demonstrate that ViDT obtains the best AP and latency trade-off among existing fully transformer-based object detectors, and achieves 49.2AP owing to its high scalability for large models. We release the code and trained models at https://github.com/naver-ai/vidt.

****

Embedding Compression with Hashing for Efficient Representation Learning in Graph

Chin-Chia Michael Yeh,Mengting Gu,Yan Zheng,Huiyuan Chen,Javid Ebrahimi,Zhongfang Zhuang,Junpeng Wang,Liang Wang,Wei Zhang

Graph neural networks (GNNs) are deep learning models designed specifically for graph data, and they typically rely on node features as the input node representation to the first layer. When applying such type of networks on graph without node feature, one can extract simple graph-based node features (e.g., number of degrees) or learn the input node representation (i.e., embeddings) when training the network. While the latter approach, which trains node embeddings, more likely leads to better performance, the number of parameters associated with the embeddings grows linearly with the number of nodes. It is therefore impractical to train the input node embeddings together with GNNs within graphics processing unit (GPU) memory in an end-to-end fashion when dealing with industrial scale graph data. Inspired by the embedding compression methods developed for natural language processing (NLP) models, we develop a node embedding compression method where each node is compactly represented with a bit vector instead of a float-point vector. The parameters utilized in the compression method can be trained together with GNNs. We show that the proposed node embedding compression method achieves superior performance compared to the alternatives.

****

BiBERT: Accurate Fully Binarized BERT

Haotong Qin,Yifu Ding,Mingyuan Zhang,Qinghua YAN,Aishan Liu,Qingqing Dang,Ziwei Liu,Xianglong Liu

The large pre-trained BERT has achieved remarkable performance on Natural Language Processing (NLP) tasks but is also computation and memory expensive. As one of the powerful compression approaches, binarization extremely reduces the computation and memory consumption by utilizing 1-bit parameters and bitwise operations. Unfortunately, the full binarization of BERT (i.e., 1-bit weight, embedding, and activation) usually suffer a significant performance drop, and there is rare study addressing this problem. In this paper, with the theoretical justification and empirical analysis, we identify that the severe performance drop can be mainly attributed to the information degradation and optimization direction mismatch respectively in the forward and backward propagation, and propose BiBERT, an accurate fully binarized BERT, to eliminate the performance bottlenecks. Specifically, BiBERT introduces an efficient Bi-Attention structure for maximizing representation information statistically and a Direction-Matching Distillation (DMD) scheme to optimize the full binarized BERT accurately. Extensive experiments show that BiBERT outperforms both the straightforward baseline and existing state-of-the-art quantized BERTs with ultra-low bit activations by convincing margins on the NLP benchmark. As the first fully binarized BERT, our method yields impressive 56.3 times and 31.2 times saving on FLOPs and model size, demonstrating the vast advantages and potential of the fully binarized BERT model in real-world resource-constrained scenarios.

****

Feature Kernel Distillation

Bobby He,Mete Ozay

Trained Neural Networks (NNs) can be viewed as data-dependent kernel machines, with predictions determined by the inner product of last-layer representations across inputs, referred to as the feature kernel. We explore the relevance of the feature kernel for Knowledge Distillation (KD), using a mechanistic understanding of an NN’s optimisation process. We extend the theoretical analysis of Allen-Zhu & Li (2020) to show that a trained NN’s feature kernel is highly dependent on its parameter initialisation, which biases different initialisations of the same architecture to learn different data attributes in a multi-view data setting. This enables us to prove that KD using only pairwise feature kernel comparisons can improve NN test accuracy in such settings, with both single & ensemble teacher models, whereas standard training without KD fails to generalise. We further use our theory to motivate practical considerations for improving student generalisation when using distillation with feature kernels, which allows us to propose a novel approach: Feature Kernel Distillation (FKD). Finally, we experimentally corroborate our theory in the image classification setting, showing that FKD is amenable to ensemble distillation, can transfer knowledge across datasets, and outperforms both vanilla KD & other feature kernel based KD baselines across a range of standard architectures & datasets.

****

Representation-Agnostic Shape Fields

Xiaoyang Huang,Jiancheng Yang,Yanjun Wang,Ziyu Chen,Linguo Li,Teng Li,Bingbing Ni,Wenjun Zhang

3D shape analysis has been widely explored in the era of deep learning. Numerous models have been developed for various 3D data representation formats, e.g., MeshCNN for meshes, PointNet for point clouds and VoxNet for voxels. In this study, we present Representation-Agnostic Shape Fields (RASF), a generalizable and computation-efficient shape embedding module for 3D deep learning. RASF is impleme nted with a learnable 3D grid with multiple channels to store local geometry. Based on RASF, shape embeddings for various 3D shape representations (point clouds, meshes and voxels) are retrieved by coordinate indexing. While there are multiple ways to optimize the learnable parameters of RASF, we provide two effective schemes among all in this paper for RASF pre-training: shape reconstruction and normal estimation. Once trained, RASF becomes a plug-and-play performance booster with negligible cost. Extensive experiments on diverse 3D representation formats, networks and applications, validate the universal effectiveness of the proposed RASF. Code and pre-trained models are publicly available\footnote{\url{https://github.com/seanywang0408/RASF}}.

****

FedProf: Selective Federated Learning with Representation Profiling

Wentai Wu,Ligang He,Weiwei Lin,carsten maple,Rui Mao

Federated Learning (FL) has shown great potential as a privacy-preserving solution to learning from decentralized data that are only accessible to end devices (i.e., clients). In many scenarios however, a large proportion of the clients are probably in possession of low-quality data that are biased, noisy or even irrelevant. As a result, they could significantly slow down the convergence of the global model we aim to build and also compromise its quality. In light of this, we propose FedProf, a novel algorithm for optimizing FL under such circumstances without breaching data privacy. The key of our approach is a data representation profiling and matching scheme that uses the global model to dynamically profile data representations and allows for low-cost, lightweight representation matching. Based on the scheme we adaptively score each client and adjust its participation probability so as to mitigate the impact of low-value clients on the training process. We have conducted extensive experiments on public datasets using various FL settings. The results show that FedProf effectively reduces the number of communication rounds and overall time (up to 4.5x speedup) for the global model to converge and provides accuracy gain.

****

Learning the Dynamics of Physical Systems from Sparse Observations with Finite Element Networks

Marten Lienen,Stephan Günnemann

We propose a new method for spatio-temporal forecasting on arbitrarily distributed points. Assuming that the observed system follows an unknown partial differential equation, we derive a continuous-time model for the dynamics of the data via the finite element method. The resulting graph neural network estimates the instantaneous effects of the unknown dynamics on each cell in a meshing of the spatial domain. Our model can incorporate prior knowledge via assumptions on the form of the unknown PDE, which induce a structural bias towards learning specific processes. Through this mechanism, we derive a transport variant of our model from the convection equation and show that it improves the transfer performance to higher-resolution meshes on sea surface temperature and gas flow forecasting against baseline models representing a selection of spatio-temporal forecasting methods. A qualitative analysis shows that our model disentangles the data dynamics into their constituent parts, which makes it uniquely interpretable.

****

Learning Synthetic Environments and Reward Networks for Reinforcement Learning

Fabio Ferreira,Thomas Nierhoff,Andreas Sälinger,Frank Hutter

We introduce Synthetic Environments (SEs) and Reward Networks (RNs), represented by neural networks, as proxy environment models for training Reinforcement Learning (RL) agents. We show that an agent, after being trained exclusively on the SE, is able to solve the corresponding real environment. While an SE acts as a full proxy to a real environment by learning about its state dynamics and rewards, an RN is a partial proxy that learns to augment or replace rewards. We use bi-level optimization to evolve SEs and RNs: the inner loop trains the RL agent, and the outer loop trains the parameters of the SE / RN via an evolution strategy. We evaluate our proposed new concept on a broad range of RL algorithms and classic control environments. In a one-to-one comparison, learning an SE proxy requires more interactions with the real environment than training agents only on the real environment. However, once such an SE has been learned, we do not need any interactions with the real environment to train new agents. Moreover, the learned SE proxies allow us to train agents with fewer interactions while maintaining the original task performance. Our empirical results suggest that SEs achieve this result by learning informed representations that bias the agents towards relevant states. Moreover, we find that these proxies are robust against hyperparameter variation and can also transfer to unseen agents.

****

Who Is Your Right Mixup Partner in Positive and Unlabeled Learning

Changchun Li,Ximing Li,Lei Feng,Jihong Ouyang

Positive and Unlabeled (PU) learning targets inducing a binary classifier from weak training datasets of positive and unlabeled instances, which arise in many real-world applications. In this paper, we propose a novel PU learning method, namely Positive and unlabeled learning with Partially Positive Mixup (P3Mix), which simultaneously benefits from data augmentation and supervision correction with a heuristic mixup technique. To be specific, we take inspiration from the directional boundary deviation phenomenon observed in our preliminary experiments, where the learned PU boundary tends to deviate from the fully supervised boundary towards the positive side. For the unlabeled instances with ambiguous predictive results, we select their mixup partners from the positive instances around the learned PU boundary, so as to transform them into augmented instances near to the boundary yet with more precise supervision. Accordingly, those augmented instances may push the learned PU boundary towards the fully supervised boundary, thereby improving the classification performance. Comprehensive experimental results demonstrate the effectiveness of the heuristic mixup technique in PU learning and show that P3Mix can consistently outperform the state-of-the-art PU learning methods.

****

Manifold Micro-Surgery with Linearly Nearly Euclidean Metrics

Jun Chen,Tianxin Huang,Wenzhou Chen,Yong Liu

The Ricci flow is a method of manifold surgery, which can trim manifolds to more regular. However, in most cases, the Rich flow tends to develop singularities and lead to divergence of the solution. In this paper, we propose linearly nearly Euclidean metrics to assist manifold micro-surgery, which means that we prove the dynamical stability and convergence of such metrics under the Ricci-DeTurck flow. From the information geometry and mirror descent points of view, we give the approximation of the steepest descent gradient flow on the linearly nearly Euclidean manifold with dynamical stability. In practice, the regular shrinking or expanding of Ricci solitons with linearly nearly Euclidean metrics will provide a geometric optimization method for the solution on a manifold.

****

Incremental False Negative Detection for Contrastive Learning

Tsai-Shien Chen,Wei-Chih Hung,Hung-Yu Tseng,Shao-Yi Chien,Ming-Hsuan Yang

Self-supervised learning has recently shown great potential in vision tasks through contrastive learning, which aims to discriminate each image, or instance, in the dataset. However, such instance-level learning ignores the semantic relationship among instances and sometimes undesirably repels the anchor from the semantically similar samples, termed as "false negatives". In this work, we show that the unfavorable effect from false negatives is more significant for the large-scale datasets with more semantic concepts. To address the issue, we propose a novel self-supervised contrastive learning framework that incrementally detects and explicitly removes the false negative samples. Specifically, following the training process, our method dynamically detects increasing high-quality false negatives considering that the encoder gradually improves and the embedding space becomes more semantically structural. Next, we discuss two strategies to explicitly remove the detected false negatives during contrastive learning. Extensive experiments show that our framework outperforms other self-supervised contrastive learning methods on multiple benchmarks in a limited resource setup.

****

Multi-Critic Actor Learning: Teaching RL Policies to Act with Style

Siddharth Mysore,George Cheng,Yunqi Zhao,Kate Saenko,Meng Wu

Using a single value function (critic) shared over multiple tasks in Actor-Critic multi-task reinforcement learning (MTRL) can result in negative interference between tasks, which can compromise learning performance. Multi-Critic Actor Learning (MultiCriticAL) proposes instead maintaining separate critics for each task being trained while training a single multi-task actor. Explicitly distinguishing between tasks also eliminates the need for critics to learn to do so and mitigates interference between task-value estimates. MultiCriticAL is tested in the context of multi-style learning, a special case of MTRL where agents are trained to behave with different distinct behavior styles, and yields up to 56% performance gains over the single-critic baselines and even successfully learns behavior styles in cases where single-critic approaches may simply fail to learn. In a simulated real-world use case, MultiCriticAL enables learning policies that smoothly transition between multiple fighting styles on an experimental build of EA’s UFC game.

****

Clean Images are Hard to Reblur: Exploiting the Ill-Posed Inverse Task for Dynamic Scene Deblurring

Seungjun Nah,Sanghyun Son,Jaerin Lee,Kyoung Mu Lee

The goal of dynamic scene deblurring is to remove the motion blur in a given image. Typical learning-based approaches implement their solutions by minimizing the L1 or L2 distance between the output and the reference sharp image. Recent attempts adopt visual recognition features in training to improve the perceptual quality. However, those features are primarily designed to capture high-level contexts rather than low-level structures such as blurriness. Instead, we propose a more direct way to make images sharper by exploiting the inverse task of deblurring, namely, reblurring. Reblurring amplifies the remaining blur to rebuild the original blur, however, a well-deblurred clean image with zero-magnitude blur is hard to reblur. Thus, we design two types of reblurring loss functions for better deblurring. The supervised reblurring loss at training stage compares the amplified blur between the deblurred and the sharp images. The self-supervised reblurring loss at inference stage inspects if noticeable blur remains in the deblurred. Our experimental results on large-scale benchmarks and real images demonstrate the effectiveness of the reblurring losses in improving the perceptual quality of the deblurred images in terms of NIQE and LPIPS scores as well as visual sharpness.

****

Learning Disentangled Representation by Exploiting Pretrained Generative Models: A Contrastive Learning View

Xuanchi Ren,Tao Yang,Yuwang Wang,Wenjun Zeng

From the intuitive notion of disentanglement, the image variations corresponding to different generative factors should be distinct from each other, and the disentangled representation should reflect those variations with separate dimensions. To discover the generative factors and learn disentangled representation, previous methods typically leverage an extra regularization term when learning to generate realistic images. However, the term usually results in a trade-off between disentanglement and generation quality. For the generative models pretrained without any disentanglement term, the generated images show semantically meaningful variations when traversing along different directions in the latent space. Based on this observation, we argue that it is possible to mitigate the trade-off by (i) leveraging the pretrained generative models with high generation quality, (ii) focusing on discovering the traversal directions as generative factors for disentangled representation learning. To achieve this, we propose Disentaglement via Contrast (DisCo) as a framework to model the variations based on the target disentangled representations, and contrast the variations to jointly discover disentangled directions and learn disentangled representations. DisCo achieves the state-of-the-art disentangled representation learning and distinct direction discovering, given pretrained non-disentangled generative models including GAN, VAE, and Flow. Source code is at https://github.com/xrenaa/DisCo.

****

Towards Building A Group-based Unsupervised Representation Disentanglement Framework

Tao Yang,Xuanchi Ren,Yuwang Wang,Wenjun Zeng,Nanning Zheng

Disentangled representation learning is one of the major goals of deep learning, and is a key step for achieving explainable and generalizable models. The key idea of the state-of-the-art VAE-based unsupervised representation disentanglement methods is to minimize the total correlation of the joint distribution of the latent variables. However, it has been proved that their goal can not be achieved without introducing other inductive biases. The Group Theory based definition of representation disentanglement mathematically connects the data transformations to the representations using the formalism of group. In this paper, built on the group-based definition and inspired by the \emph{n-th dihedral group}, we first propose a theoretical framework towards achieving unsupervised representation disentanglement. We then propose a model based on existing VAE-based methods to tackle the unsupervised learning problem of the framework. In the theoretical framework, we prove three sufficient conditions on model, group structure, and data respectively in an effort to achieve, in an unsupervised way, disentangled representation per group-based definition. With these conditions, we offer an option, from the perspective of the group-based definition, for the inductive bias that existing VAE-based models lack. Experimentally, we train 1800 models covering the most prominent VAE-based methods on five datasets to verify the effectiveness of our theoretical framework. Compared to the original VAE-based methods, these Groupified VAEs consistently achieve better mean performance with smaller variances.

****

Learning Efficient Image Super-Resolution Networks via Structure-Regularized Pruning

Yulun Zhang,Huan Wang,Can Qin,Yun Fu

Several image super-resolution (SR) networks have been proposed of late for efficient SR, achieving promising results. However, they are still not lightweight enough and neglect to be extended to larger networks. At the same time, model compression techniques, like neural architecture search and knowledge distillation, typically consume considerable computation resources. In contrast, network pruning is a cheap and effective model compression technique. However, it is hard to be applied to SR networks directly because filter pruning for residual blocks is well-known tricky. To address the above issues, we propose structure-regularized pruning (SRP), which imposes regularization on the pruned structure to ensure the locations of pruned filters are aligned across different layers. Specifically, for the layers connected by the same residual, we select the filters of the same indices as unimportant filters. To transfer the expressive power in the unimportant filters to the rest of the network, we employ L_2 regularization to drive the weights towards zero so that eventually, their absence will cause minimal performance degradation. We apply SRP to train efficient image SR networks, resulting in a lightweight network SRPN-Lite and a very deep one SRPN. We conduct extensive comparisons with both lightweight and larger networks. SRPN-Lite and SRPN perform favorably against other recent efficient SR approaches quantitatively and visually.

****

Imitation Learning from Pixel Observations for Continuous Control

Samuel Cohen,Brandon Amos,Marc Peter Deisenroth,Mikael Henaff,Eugene Vinitsky,Denis Yarats

We study imitation learning using only visual observations for controlling dynamical systems with continuous states and actions. This setting is attractive due to the large amount of video data available from which agents could learn from. However, it is challenging due to $i)$ not observing the actions and $ii)$ the high-dimensional visual space. In this setting, we explore recipes for imitation learning based on adversarial learning and optimal transport. A key feature of our methods is to use representations from the RL encoder to compute imitation rewards. These recipes enable us to scale these methods to attain expert-level performance on visual continuous control tasks in the DeepMind control suite. We in vestigate the tradeoffs of these approaches and present a comprehensive evaluation of the key design choices. To encourage reproducible research in this area, we provide an easy-to-use implementation for benchmarking visual imitation learning, including our methods.

****

Probabilistic Implicit Scene Completion

Dongsu Zhang,Changwoon Choi,Inbum Park,Young Min Kim

We propose a probabilistic shape completion method extended to the continuous geometry of large-scale 3D scenes. Real-world scans of 3D scenes suffer from a considerable amount of missing data cluttered with unsegmented objects. The problem of shape completion is inherently ill-posed, and high-quality result requires scalable solutions that consider multiple possible outcomes. We employ the Generative Cellular Automata that learns the multi-modal distribution and transform the formulation to process large-scale continuous geometry. The local continuous shape is incrementally generated as a sparse voxel embedding, which contains the latent code for each occupied cell. We formally derive that our training objective for the sparse voxel embedding maximizes the variational lower bound of the complete shape distribution and therefore our progressive generation constitutes a valid generative model. Experiments show that our model successfully generates diverse plausible scenes faithful to the input, especially when the input suffers from a significant amount of missing data. We also demonstrate that our approach outperforms deterministic models even in less ambiguous cases with a small amount of missing data, which infers that probabilistic formulation is crucial for high-quality geometry completion on input scans exhibiting any levels of completeness.

****

Improving Federated Learning Face Recognition via Privacy-Agnostic Clusters

Qiang Meng,Feng Zhou,Hainan Ren,Tianshu Feng,Guochao Liu,Yuanqing Lin

The growing public concerns on data privacy in face recognition can be partly relieved by the federated learning (FL) paradigm. However, conventional FL methods usually perform poorly due to the particularity of the task, \textit{i.e.}, broadcasting class centers among clients is essential for recognition performances but leads to privacy leakage. To resolve the privacy-utility paradox, this work proposes PrivacyFace, a framework largely improves the federated learning face recognition via communicating auxiliary and privacy-agnostic information among clients. PrivacyFace mainly consists of two components: First, a practical Differentially Private Local Clustering (DPLC) mechanism is proposed to distill sanitized clusters from local class centers. Second, a consensus-aware recognition loss subsequently encourages global consensuses among clients, which ergo leads to more discriminative features. The proposed schemes are mathematically proved to be differential private, introduce a lightweight overhead as well as yield prominent performance boosts (\textit{e.g.}, +9.63\% and +10.26\% for TAR@FAR=1e-4 on IJB-B and IJB-C respectively). Extensive experiments and ablation studies on a large-scale dataset have demonstrated the efficacy and practicability of our method.

****